MedMirror: Towards More Reliable Diagnosis in Traditional Chinese Medicine via Reflexive Interaction and Multi-Agent Collaboration

Anonymous ACL submission

Abstract

Despite recent advancements in specialized Traditional Chinese Medicine (TCM) AI systems, they remain constrained by modal inflexibility, lack of reliability, and insufficient explainability. To address these issues, we propose **MedMirror**, a framework that focuses on two key components of medical consultation AI systems: diagnostic inquiry and evidence-based explanation. For the diagnostic inquiry phase, we propose User-Centric Reflexive Diagnostic Interaction. It leverages dual agents to perform dynamic, multi-turn inquiries, ensuring that comprehensive evidence is gathered to facilitate highly reliable diagnosis. For the evidence-based explanation phase, we propose Multi-Agent Collaborative Knowledge Synthesis & Report Generation. This module synthesizes diagnostic data through multi-path parallel RAG, reflexive argumentation, and iterative drafting. It effectively transforms clinical findings into comprehensive, accessible, and evidence-backed reports for users. Experimental results demonstrate that MedMirror achieves superior performance in syndrome differentiation compared to existing baselines. Notably, its reflexive mechanism effectively mitigates information deficiency, while expert meta-evaluation confirms the system's effectiveness in producing high-quality and reliable diagnostic insights.

1 Introduction

Recent advancements in Large Language Models (LLMs) (Meta et al., 2024; Hurst et al., 2024; OpenAI et al., 2024; DeepSeek-AI et al., 2025) have laid a solid foundation for applications in specialized domains. Traditional Chinese Medicine (TCM) represents a particularly crucial field due to its unique theoretical system and clinical value (Zou et al., 2023; Lu et al., 2004; Ma et al., 2019). Domain-specific LLMs such as HuaTuo (Wang et al., 2023), ShenNong (Wei Zhu and Wang, 2023), Zhongjing (Yang et al., 2023), SunSimiao (Lab, 2023), and the agent-based JingFang (Yang et al., 2025) have established a preliminary framework for tasks like syndrome differentiation[1].

However, despite these strides, applying general LLM paradigms to healthcare reveals a critical structural mismatch. Real-world patient queries are inherently sparse and fragmented. Current AI systems predominantly operate as "passive responders", forcing immediate outputs based on users' inputs, no matter weather it is complete. This passive instruction-following inevitably leads to premature diagnostic inferences and hallucinations, as the model lacks the agency to verify missing evidence. Furthermore, the "black-box" nature of end-to-end generation fails to provide the transparent evidence chains required for clinical trust.

In medical domains, the cost of a diagnostic error far outweighs the computational cost of additional inference tokens. Unlike general chatbots optimized for latency, MedMirror prioritizes clinical rigor and safety. We argue that a reliable medical AI must transition from a "fast thinking" (direct generation) paradigm to a "slow thinking" (reflective and iterative) architecture.

From a user's perspective, an ideal consultation experience relies on comprehensive and purposeful interaction that allows the clinician to fully understand the clinical context. Beyond providing a diagnosis, it is essential to offer the underlying rationale to foster patient trust and ensure reliability.

To address these limitations and replicate the ideal experience, we propose MedMirror. Specifically, our work focuses on two primary dimensions: the external diagnostic inquiry mechanism and internal evidence-based report generation. For the diagnostic inquiry phase, we introduce User-Centric Reflexive Diagnostic Interaction based on two components: (1) a multimodal framework that incorporates a Tongue Diagnosis Agent for cross-

[1] See Appendix B for detailed related work.

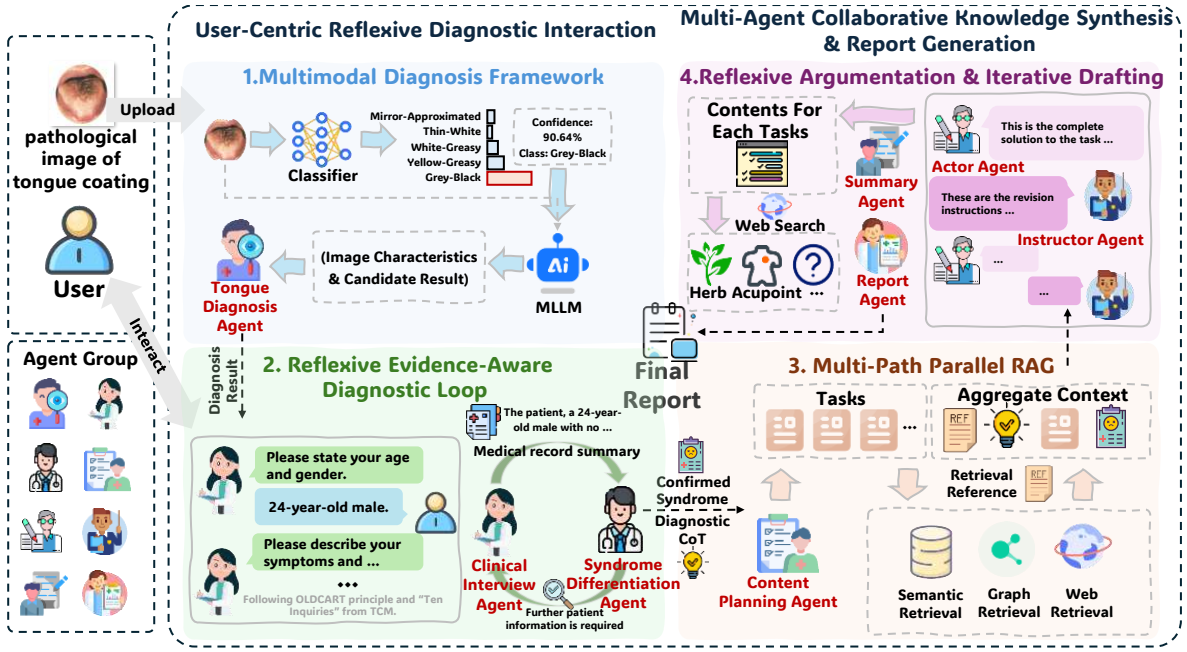

Figure 1: Overview of the MedMirror framework. The system incorporates eight agents organized into two main parts, with each part further divided into two modules. Users can provide multimodal inputs, including tongue image, and engage in multi-turn dialogues to obtain final diagnostic reports.

modal perception, and (2) the Reflexive Evidence-Aware Diagnostic Loop (READ-Loop), a credible interaction mechanism that dynamically integrates inquiry with diagnosis. This ensures that the entire reasoning process is transparent, traceable, and supported by sufficient evidence. For the diagnostic reporting phase, we develop Multi-Agent Collaborative Knowledge Synthesis & Report Generation. To bridge the gap between diagnostic conclusions and actionable advice, we design a pipeline integrating Content Planning, Multi-Path Parallel Retrieval-Augmented Generation (MPPR), and Reflexive Argumentation & Iterative Drafting (RAID). This internal collaboration transforms raw diagnostic data into comprehensive, explainable reports enriched with multimodal elements and personalized guidance.

To evaluate the diagnostic effectiveness of MedMirror, we use TCM-SD (Ren et al., 2022) and TCMEval-SDT (Cheng et al., 2025) for fine-grained syndrome differentiation. We further validate the system through multimodal assessments on a tongue diagnosis dataset (PaddlePaddleAIStudio, 2021), interactive capability tests under data-sparse conditions, and robustness analyses including ablation and scalability studies. Finally, to ensure clinical validity, we conduct a qualitative assessment comprising detailed case studies and expert-driven meta-evaluations of diagnostic report quality.

2 Methodology

2.1 Problem Analysis & Design Concepts

Problem Analysis. A significant paradox exists in medical consultation AI systems: while LLMs often outmatch junior physicians in knowledge reserves, their practical implementation remains limited. This stems from two primary factors. First, user-provided symptom descriptions are typically characterized by informational sparsity, lacking the clinical rigor required for accurate assessment. Current systems tend to be passively following user prompts and forcing premature diagnostic inferences without sufficient evidence. This lack of proactive verification inevitably results in model hallucinations. Second, in safety-critical domains such as healthcare, providing isolated results without a clear diagnostic rationale fails to establish user trust. Furthermore, reports dense with clinical jargon often remain inaccessible to lay users. Consequently, generating comprehensive, evidence-backed reports that are also human-readable is essential for practical clinical utility.

Design Concepts. Our entire workflow is designed to emulate real-world clinical consultations, prioritizing both diagnostic accuracy and user experience. We construct a multi-agent system with dedicated workflows tailored for each stage of medical inquiry, diagnosis, and report generation. Notably, while

MEDMIRROR is developed for TCM, its underlying methodology and architectural paradigm are generalizable to other clinical domains, as empirically validated through our experiments in Section 3.3.4.

2.2 System Overview

To resolve the forementioned information asymmetry paradox, MedMirror establishes a "dual-loop" collaborative framework (see Fig. 1)[2]. Unlike conventional pipelines that passively process static inputs, our system orchestrates two dynamic phases designed to mirror the workflow of expert clinicians: Active Evidence Acquisition and Rigorous Knowledge Synthesis.

Phase I: Bridging the Information Gap (External Loop). Addressing the challenge of informational sparsity, the first phase deploys the User-Centric Reflexive Diagnostic Interaction. Here, the system does not merely "accept" user input; instead, it initiates a Reflexive Evidence-Aware Diagnostic Loop (READ-Loop). By acting as an active inquirer rather than a passive responder, the system dynamically updates the medical record summary through multi-turn dialogue, effectively "forcing" the completion of the diagnostic evidence chain before any conclusion is drawn.

Phase II: From Diagnosis to Understanding (Internal Loop). A diagnosis is only as effective as the patient's understanding of it. To address the challenges of "black-box" outputs and inaccessible jargon, this phase executes Multi-Agent Collaborative Knowledge Synthesis. Instead of simply outputting a raw diagnostic label, the system employs Reflexive Argumentation & Iterative Drafting (RAID) to act as a rigorous "medical communicator". Similar to how a conscientious doctor drafts a discharge summary, agents in this loop iteratively retrieve authoritative evidence (via MPPR) to substantiate the diagnosis, "translate" specialized TCM terminology into readable text, and refine the report. This ensures the final output is not just clinically accurate, but also transparent, evidence-backed, and fully accessible to lay users.

2.3 User-Centric Reflexive Diagnostic Interaction

This phase emulates proactive consultation via a "Perceive-then-Inquire" workflow to mitigate information sparsity. First, the Multimodal Diagnosis Framework extracts visual evidence as a diagnostic prior. Subsequently, the READ-Loop iteratively identifies information gaps and generates targeted questions to resolve ambiguities. This collaboration ensures the final diagnosis is supported by a complete and consistent evidence chain.

2.3.1 Multimodal Diagnosis Framework

The Tongue Diagnosis Agent (TDA) functions as a composite mapping $\mathcal{F}_{\text{TDA}} : \mathbf{x}_{\text{img}} \to s_{\text{diag}}$, transforming the raw tongue image $\mathbf{x}_{\text{img}}$ into a clinical summary through a "discriminate-describe-synthesiz" pipeline. First, a classifier $\mathcal{C}$ extracts a robust categorical prior $c_{\text{prior}} = \mathcal{C}(\mathbf{x}_{\text{img}})$ to enforce pathological constraints. Conditioned on this prior, a Multimodal LLM $\mathcal{M}$ generates fine-grained visual descriptions $d_{\text{desc}} = \mathcal{M}(\mathbf{x}_{\text{img}}, c_{\text{prior}})$, capturing subtle features like texture or cracks. Finally, an LLM $\mathcal{L}$ synthesizes these signals to derive the diagnostic conclusion $s_{\text{diag}} = \mathcal{L}(c_{\text{prior}}, d_{\text{desc}})$. This hierarchical design effectively anchors the interpretability of generative models with the objective precision of discriminative classifiers.

2.3.2 Reflexive Evidence-Aware Diagnostic Loop (READ-Loop)

To reduce hallucinations caused by passive generation from sparse inputs, READ-Loop models diagnosis as a dynamic state-refinement process to maximize evidence sufficiency. Let S_t denote the medical record summary at turn t. The system orchestrates a Clinical Interview Agent (A_{CIA}) for information seeking and a Syndrome Differentiation Agent (A_{SDA}) for reasoning.

At each step t, A_{SDA} assesses S_t by accessing three external knowledge modules: a case vector index $\mathcal{K}_{\text{vec}}$, a syndrome taxonomy $\mathcal{K}_{\text{tax}}$, and a knowledge graph $\mathcal{G}$. We define the reasoning process as a composite mapping $\Phi : S_t \to (\hat{y}, \mathcal{M}_{\text{miss}})$, executing a three-stage reasoning protocol:

1. Case-Driven Retrieval: Retrieve top-k similar cases from $\mathcal{K}_{\text{vec}}$ to establish a diagnostic prior $P(y|S_t)$.
2. Taxonomy-Guided Refinement: Filter candidates through $\mathcal{K}_{\text{tax}}$ to enforce logical constraints and cluster related syndromes.
3. Graph-Based Verification: Ground the refined hypothesis in $\mathcal{G}$ to verify path consistency.

Crucially, the loop incorporates a decision gate based on the missing information set $\mathcal{M}_{\text{miss}}$. If the evidence is insufficient (i.e., $\mathcal{M}_{\text{miss}} \neq \emptyset$), the system triggers the Inquiry Phase: A_{CIA} generates

[2] See Appendix F for the detailed pseudocode of each component.

a targeted question q_t based on $\mathcal{M}_{\text{miss}}$ with the OLDCART [3] and TCM Ten-Inquiry [4] schemas to acquire user feedback. The state is then updated as $S_{t+1} = S_t \oplus \{q_t, r_t\}$. This cycle repeats until $\mathcal{M}_{\text{miss}} = \emptyset$ or a maximum turn T is reached, ensuring the final diagnosis $\hat{y}$ is derived from a converged, evidence-complete state rather than premature speculation.

2.4 Multi-Agent Collaborative Knowledge Synthesis & Report Generation

Upon determining the diagnosis, the system transitions from inquiry to explanation. To bridge the gap between raw clinical reasoning and patient accessibility, we orchestrate a hierarchical "Plan-Retrieve-Refine" workflow. As the central orchestrator, the Content Planning Agent (CPA) first decomposes the diagnosis into discrete writing tasks. These tasks then drive MPPR to retrieve targeted authoritative evidence, which feeds into the RAID mechanism for iterative drafting and verification. This collaboration ensures the final report is not only clinically accurate but also transparent and user-friendly.

2.4.1 Content Planning Agent (CPA)

As the central orchestrator, the CPA translates diagnostic inputs into a structured sequence of writing tasks. This formalized task list directs the RAID module for section-specific generation and drives the MPPR module to execute targeted reference retrieval. By decomposing the report into discrete units, CPA effectively bridges the gap between diagnostic reasoning and systematic content synthesis.

2.4.2 Multi-Path Parallel RAG (MPPR)

MPPR constructs a comprehensive evidence space $\mathcal{E}$ for the drafting agents by aggregating two complementary layers: Cross-Task Shared References ($\mathcal{E}_{\text{sha}}$) and Task-Specific References ($\mathcal{E}_{\text{spec}}$). This dual-layer architecture balances broad clinical consensus with granular, task-relevant details.

Cross-Task Shared References. This layer provides a static context accessible to all tasks, derived from the confirmed syndrome s_{diag}. We define $\mathcal{E}_{\text{sha}}$ as the union of three retrieval streams:

$$\mathcal{E}_{\text{sha}} = \mathcal{R}_{\text{vec}}(s_{\text{diag}}) \cup \mathcal{R}_{\text{web}}(s_{\text{diag}}) \cup \mathcal{R}_{\text{kg}}(s_{\text{diag}}). \quad (1)$$

Here, $\mathcal{R}_{\text{vec}}$ retrieves semantically relevant passages from TCM classics via vector similarity (filtered by an LLM for pertinence), $\mathcal{R}_{\text{web}}$ captures broader clinical consensus through general web queries, and $\mathcal{R}_{\text{kg}}$ extracts structured syndrome attributes (e.g., manifestations, treatment principles) directly from the internal Knowledge Graph.

Task-Specific References. Unlike the shared layer, this set is dynamically instantiated for each specific writing task t_i. The mechanism generates a specialized query q_i based on the task requirements to retrieve focused evidence: $\mathcal{E}^{(i)}_{\text{spec}} = \mathcal{R}_{\text{web}}(q_i)$. This ensures that drafting agents possess precise information necessary for composing distinct report sections (e.g., retrieving contraindications for a specific herb mentioned in the prescription).

2.4.3 Reflexive Argumentation & Iterative Drafting (RAID)

To ensure clinical rigor, we formalize the drafting process as a cooperative optimization game between an Actor Agent (A_{act}) and an Instructor Agent (A_{ins}). Let $\mathbf{x} = (\mathcal{E}_{\text{spec}}, s_{\text{diag}}, \text{CoT})$ denote the input context containing retrieved evidence, diagnostic state, and reasoning chains. The process initializes with A_{act} generating a draft $\mathcal{O}_0 = A_{\text{act}}(t_i, \mathbf{x})$ for task t_i.

The system executes a reflexive loop to optimize $\mathcal{O}$. At each step k, A_{ins} functions as a critic $V(\mathcal{C}_k) \to \{\text{pass}, \mathcal{I}_{\text{fb}}\}$, evaluating the draft against a criteria set Φ: factual accuracy, task adherence, and citation fidelity. If criteria are not met, A_{ins} generates specific refinement instructions $\mathcal{I}_{\text{fb}}$, prompting an update:

$$\mathcal{O}_{k+1} = A_{\text{act}}(\mathcal{O}_k, \mathcal{I}_{\text{fb}}, \mathbf{x}). \quad (2)$$

This cycle repeats until $V(\mathcal{O}_k) = \text{pass}$ or a maximum step count is reached. Finally, a Summary Agent aggregates the optimization trajectory to produce the coherent section report. This modular architecture allows all tasks $\{t_i\}$ assigned by the CPA to be processed in parallel, significantly reducing global latency.

2.4.4 Report Agent

The Report Agent functions as the final synthesis engine, converting the textual draft into a user-friendly multimodal narrative. It first parses key clinical entities (e.g., herbs, acupoints) and augments them with retrieved visual assets and authoritative explanatory links. Furthermore, the agent integrates personalized lifestyle guidance via web search. This fusion ensures the final report balances clinical rigor with visual accessibility and actionable advice.

[3] OLDCART: Onset, Location, Duration, Characteristics, Aggravating factors, Relieving factors, and Treatment.

[4] TCM Ten-Inquiry is a traditional Chinese medicine diagnostic schema for systematic symptom gathering.

3 Experiments

We evaluate MedMirror through a multi-tiered framework, beginning with fine-grained syndrome differentiation on TCM-SD (Ren et al., 2022) and TCMEval-SDT (Cheng et al., 2025), supplemented by multimodal tongue diagnosis (PaddlePaddleAIStudio, 2021) and interactive information mining. System robustness is further verified through component ablation, scalability testing, and quantitative output quality evaluation, alongside a cross-domain evaluation assessing adaptability to western medicine diagnostics. Finally, clinical validity is substantiated via qualitative case studies and expert-driven meta-evaluations of evidence chain completeness, with details provided in Appendix C. For further details, refer to Appendix D.2 for a complete case study, Appendix G for the system prompts, and Appendix H for samples of the final diagnostic reports.

3.1 Core Diagnostic Performance

3.1.1 Syndrome Differentiation

As shown in Table 1, MedMirror consistently achieves superior performance across all benchmarks. The Kimi-K2 variant reaches SOTA on TCMEval-SDT (Accuracy: 55.00%), while the DeepSeek-V3 variant leads on TCM-SD (Weighted F1: 0.7063).The RAG w/ CoT baseline validates the necessity of external knowledge by boosting DeepSeek-V3's performance (Weighted F1: 0.3221 to 0.6193). However, this approach remains substantially behind MedMirror (Weighted F1: 0.7063). This performance gap in Weighted F1 demonstrates that passive retrieval is insufficient, as MedMirror's reflexive verification is essential to filter noise and enforce pathological consistency. While traditional multi-agent strategies such as Voting and Debate offer marginal gains, they fail to match the efficacy of knowledge-augmented approaches. Notably, MedMirror enables the smaller HuaTuo-o1-7B to outperform the unenhanced DeepSeek-V3 Base (Accuracy: 42.00% Accuracy: vs. 33.00%), which proves that structured domain reasoning contributes more to diagnostic precision than model scale alone.

3.1.2 Tongue Diagnosis

Extensive benchmarking of convolutional architectures (detailed in Appendix D.1 Table 6) identifies DenseNet-201 (Huang et al., 2018) as the optimal visual encoder, achieving a superior Macro F1-score of 0.9036. This selection is critical not merely for classification accuracy, but for establishing a robust diagnostic prior. By anchoring the subsequent multimodal reasoning in high-fidelity categorical predictions, this module effectively creates a visual guardrail, mitigating the risk of hallucinatory descriptions often observed in pure generative VLMs when processing subtle medical textures (e.g., distinguishing *greasy* vs. *rotting* coatings).

3.2 Interactive Capability

We evaluate the evidence-gathering capability of the READ-Loop under information-deficient conditions. The experiment simulate incomplete diagnostic inputs by providing only brief chief complaint, omitting other detailed information. An LLM agent acted as a patient proxy, responding to inquiries based on ground-truth clinical data.

Fig. 2 reveals that while iterative inquiry effectively enhances diagnostic performance, scaling trajectories are highly backbone-dependent and non-monotonic. DeepSeek-V3 demonstrates a rapid initial surge (F1: $0.3401 \rightarrow$ F1: 0.5797) but suffers from premature degradation, suggesting susceptibility to semantic drift as dialogue context expands. In contrast, Huatuo-o1-7B and Kimi-K2 exhibit greater resilience to prolonged inquiry; the former surpasses the full-information baseline (0.5633) by the second iteration, while the latter achieves the highest diagnostic ceiling (F1: 0.6214). Ultimately, these results underscore a critical trade-off: although iterative probing compensates for initial information scarcity, its efficacy is bounded by cumulative noise and the backbone's inherent reasoning stability, necessitating a synergy between sophisticated reasoning and comprehensive initial clinical evidence.

3.3 Analysis and Robustness

3.3.1 Ablation Study

Ablation results in Table 2 delineate the contributions of MedMirror's core modules. The Medical Cases Database is the most critical component; its removal precipitated substantial degradation (F1: -34.21%; Recall: -37.30%), underscoring the necessity of case-based grounding. The Knowledge Graph and Taxonomy Table provide essential structural support, with their exclusion leading to moderate declines in Recall and F1, respectively. Ultimately, the full configuration achieves peak performance (F1: 0.7063), confirming that the synergistic integration of retrieval, refinement, and

Table 1: Performance comparison on TCMEval-SDT and TCM-SD benchmarks. Darkest blue, medium blue, and light blue indicate the 1st, 2nd, and 3rd best performance respectively. n: agents, r: debate rounds.

Kernel Model		TCMEval-SDT				TCM-SD		
		P_{avg}	R_{avg}	$F1_{avg}$	**Acc**	P_{wei}	R_{wei}	$F1_{wei}$
MedMirror	DeepSeek-V3 (DeepSeek-AI et al., 2025)	0.6750	0.6817	0.6743	49.00%	**0.7271**	**0.7130**	**0.7063**
	Kimi-K2 (Team et al., 2025b)	**0.7200**	0.7067	0.7083	**55.00%**	0.7195	0.6778	0.6806
	HuaTuo-o1-7B (Wang et al., 2023)	0.6100	0.6033	0.6000	42.00%	0.6039	0.5660	0.5633
RAG w/ CoT	DeepSeek-V3 (DeepSeek-AI et al., 2025)	0.6983	0.7450	0.6913	36.00%	0.6290	0.6500	0.6193
	Kimi-K2 (Team et al., 2025b)	0.7100	0.7633	**0.7117**	40.00%	0.6196	0.4600	0.5032
	HuaTuo-o1-7B (Wang et al., 2023)	0.6200	0.5500	0.5550	30.00%	0.6505	0.3500	0.3997
Vote ($n = 3$)	DeepSeek-V3 (DeepSeek-AI et al., 2025)	0.7017	0.7533	0.7017	38.00%	0.5284	0.3690	0.4119
	Kimi-K2 (Team et al., 2025b)	0.7050	0.7000	0.6740	37.00%	0.4919	0.3160	0.3626
	HuaTuo-o1-7B (Wang et al., 2023)	0.6183	0.5933	0.5713	29.00%	0.3678	0.2619	0.2621
Debate ($n = 3, r = 2$)	DeepSeek-V3 (DeepSeek-AI et al., 2025)	0.6633	0.7683	0.6857	33.00%	0.5468	0.3650	0.4151
	Kimi-K2 (Team et al., 2025b)	0.6800	0.7667	0.6950	38.00%	0.4862	0.3215	0.3636
	HuaTuo-o1-7B (Wang et al., 2023)	0.5750	0.6300	0.5710	23.00%	0.4064	0.2900	0.2883
Base LLMs w/ CoT	DeepSeek-V3 (DeepSeek-AI et al., 2025)	0.6617	**0.7733**	0.6900	33.00%	0.4700	0.3032	0.3221
	Kimi-K2 (Team et al., 2025b)	0.6950	0.7317	0.6907	39.00%	0.4678	0.2999	0.3394
	GPT-4o (Hurst et al., 2024)	0.5500	0.6883	0.5900	24.00%	0.4328	0.2138	0.2474
	Gemini-2.0 (Team et al., 2025a)	0.4965	0.6767	0.5500	17.00%	0.4420	0.2480	0.2810
	Claude-3.7 (Anthropic, 2025)	0.5800	0.7117	0.6133	23.00%	0.2731	0.1206	0.1453
	HuaTuo-o1-7B (Wang et al., 2023)	0.5600	0.5450	0.5280	24.00%	0.3339	0.2408	0.2473
	Sunsimiao-7B (Lab, 2023)	0.4167	0.4250	0.4013	18.00%	0.2122	0.0660	0.0879
	Carebot-8B (Zhao et al., 2024)	0.2723	0.2350	0.2267	10.00%	0.2596	0.1380	0.1369

verification is indispensable for addressing the complexity of syndrome differentiation.

3.3.2 Scalability Analysis

Scalability evaluations on the Qwen2.5 series (Qwen et al., 2025) (see Fig. 3) demonstrate that MedMirror consistently enhances performance across various parameter scales. While the 7B model shows a modest gain (+1%), the 14B variant exhibits a 60.0% relative increase from 20% to 32%. The 32B model achieves peak accuracy at 42% (+35.5% relative). Notably, MedMirror rectifies the performance degradation observed in the 72B base model, providing a +15% absolute improvement (+62.5% relative). These findings verify that the framework is highly scalable and effectively unlocks latent domain knowledge in large-scale models.

3.3.3 Output Quality Evaluation

We assess generative quality using DeepSeek-R1 (DeepSeek-AI et al., 2025) as an impartial judge. As shown in Table 3, the RAID mechanism shifts generation from unconstrained fluency toward evidence-based precision, increasing Fidelity to 0.8149 and reducing Conflict Scores to 0.1369. Although this filtering leads to a marginal decline in Clinical Sufficiency (3.76 vs. 3.97), Citation Accuracy improves to 4.28. This trade-off demonstrates that RAID enforces "conservative rigor," prioritizing clinical safety and verifiable attribution over the unverified content characteristic of the baseline.

3.3.4 Cross-Domain Evaluation

To verify the architectural universality of MedMirror beyond TCM, we evaluated it on the Western Medicine CMB benchmark. As shown in Table 4, the model demonstrates strong adaptability in standard tasks, elevating MCQ accuracy from 0.8727 to 0.9158. A critical insight emerges from the Clinical QA results. In the "Full Description" setting—where the input represents a perfect, doctor-curated medical record—MedMirror exhibits a performance regression compared to the baseline (Accuracy: 0.7968 vs. Accuracy: 0.8472). We attribute this to information saturation: when the context is already complete, the READ-Loop's forced active inquiry may introduce conversational noise or hallucinated details, disrupting the base model's direct inference path. However, we argue that "perfect input" is an artificial construct. In real-world telemedicine, patients rarely provide structured, exhaustive medical histories. They present with fragmented complaints (e.g., "my stomach

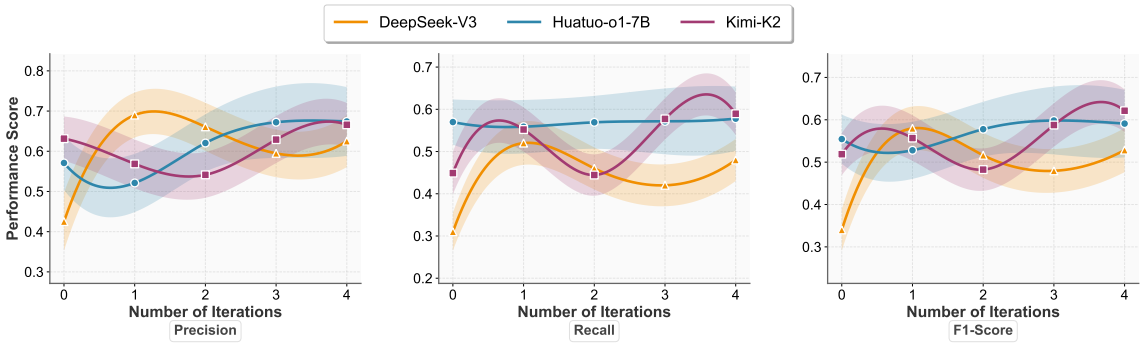

Figure 2: Performance of READ-Loop under insufficient information with different loop counts on the TCM-SD.

hurts"). Under this "Only Chief Complaint" setting—which mirrors ecological clinical reality—the baseline collapses (Accuracy: 0.3802), exposing the fragility of passive generation. In stark contrast, with only one single iteration of the READ-Loop, MedMirror systematically bridges this evidentiary gap through active inquiry, recovering the performance to 0.6833 (Accuracy: +30.31%). Conclusion. This contrast highlights the core value of MedMirror: while it may incur a minor trade-off in idealized, information-rich scenarios, it offers critical resilience in the ubiquitous, high-uncertainty scenarios of actual patient interaction. It shifts the system from a "text processor" to a "clinical investigator".

3.3.5 Cost-Efficiency Analysis

Fig. 4 illustrates a strategic trade-off between computational cost and diagnostic precision. Despite a substantial token overhead (10.9× vs. Base LLM; +159% vs. RAG), MedMirror yields significant performance dividends: it achieves an F_1 of 0.7063 in sufficient-info scenarios, outperforming both RAG (0.6193) and the base model (0.3221). Critically, in sparse-info settings where RAG's gains are marginal (F_1: 0.3282), MedMirror (It=1) prevents model collapse, reaching an F_1 of 0.5797. These results suggest that while RAG offers cost-effective retrieval, it lacks the robust reasoning required for insufficient clinical data. MedMirror marks a paradigm shift from "fast-thinking" retrieval to a "slow-thinking" reflective architecture, prioritizing diagnostic reliability over inference latency—a necessary prioritization given the high clinical stakes of misdiagnosis.

3.4 Meta-Evaluation

We conduct a meta-evaluation of the system-generated reports, with further details available in Appendix C.5. Table 5 details the expert evaluation results, where MedMirror achieves a superior total score of 22.80 ± 1.37, significantly outperforming the ablation baseline without RAID (18.00 ± 2.35). This substantial disparity is most pronounced in Evidence Chain Completeness (4.73 vs. 3.00) and Tongue Image Utilization (4.53 vs. 3.20), underscoring the critical role of the reflexive mechanism in ensuring logical traceability and effectively anchoring multimodal evidence. Furthermore, the significantly higher variance observed in the baseline (± 2.35) indicates that without iterative refinement,

Table 2: Ablation study on the Syndrome Differentiation Agent's performance with different configurations on TCM-SD.

Configuration	P	R	F1
w/o KG	**0.7395** (+1.24%)	0.6590 (-5.40%)	0.6811 (-2.52%)
w/o DB	0.5254 (-20.17%)	0.3400 (-37.30%)	0.3642 (-34.21%)
w/o TT	0.7242 (-0.29%)	0.6503 (-6.27%)	0.6688 (-3.75%)
MedMirror	0.7271	**0.7130**	**0.7063**

*Note:*DB denotes the Medical Cases Database; KG refers to the Syndrome Knowledge Graph; TT represents the Syndrome Taxonomy Table.

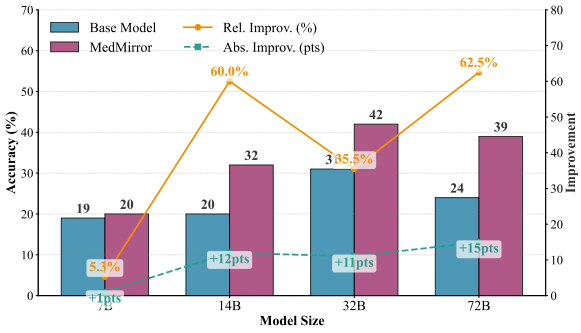

Figure 3: Scalability analysis of MedMirror versus the Qwen 2.5 base models on TCMEval-SDT.

Table 3: Comparison of output quality evaluations.

Metric	MPPR+Single	MPPR+RAID
Fact-based Consistency		
Fid. Score (↑)	0.6296	**0.8149**
Conf. Score (↓)	0.3561	**0.1369**
Quality Scoring (1–5 Scale)		
Clin. Suff. (↑)	**3.9662**	3.7568
Task Comp. (↑)	**4.9932**	4.9122
Cite. Acc. (↑)	3.6959	**4.2770**

Note: ↑ (↓) indicates higher (lower) is better. **Bold** indicates superior performance ($p < 0.05$ or best).

Table 4: Cross-domain evaluation results on the CMB benchmark.

Task	Kernel Model	Accuracy
MCQ	DeepSeek-V3	0.8727
	MedMirror	**0.9158**
Clinical QA (Full Description)	DeepSeek-V3	**0.8472**
	MedMirror	0.7968
Clinical QA (Only Chief Complaint)	DeepSeek-V3	0.3802
	MedMirror (READ-Loop iter=1)	**0.6833**

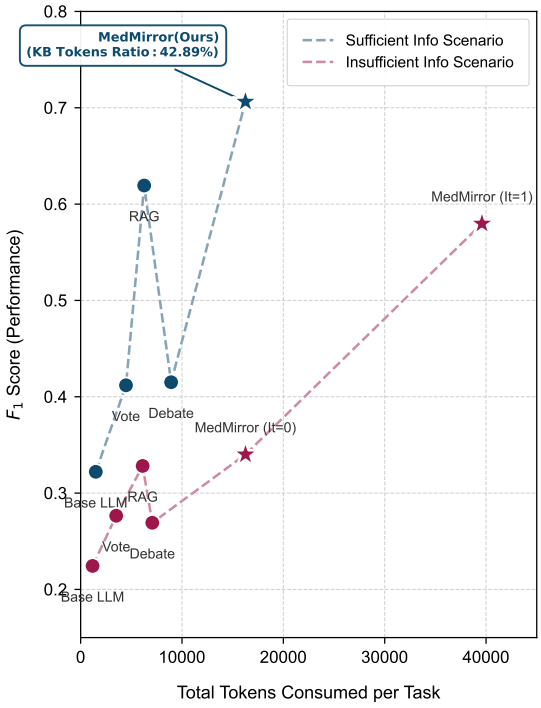

Figure 4: Cost-Efficiency analysis of MedMirror. KB Tokens Ratio represents the average proportion of tokens allocated to the MedMirror external knowledge base per reasoning task.

the system suffers from instability in diagnostic reasoning. While the baseline remains relatively competitive in textual Explainability (DCE: 4.40), its failure to maintain high standards in clinical rigor (ECC and SDS) confirms that the RAID module effectively shifts the generation paradigm from superficial fluency to verifiable clinical reliability.

4 Conclusion

In this work, we present MedMirror, a framework that fundamentally shifts medical AI from passive information processing to active clinical investigation. By orchestrating the reflexive READ-Loop and the rigorous RAID mechanism, our system systematically resolves the ambiguity inherent in sparse patient queries. Crucially, our architecture embodies a "Safety over Efficiency" philosophy: we demonstrate that trading computational overhead for iterative verification significantly mitigates hallucinations and establishes a complete, traceable evidence chain. Beyond TCM, the proven adaptability of MedMirror to Western medicine suggests a scalable path toward trustworthy, evidence-grounded decision support systems in high-stakes domains.

Table 5: Comparison of Average Meta-Evaluation scores for MedMirror's diagnostic reports.

Evaluation Dimension	MedMirror (Mean ± SD)	w/o RAID (Mean ± SD)
ECC	4.73 ± 0.46	3.00 ± 0.71
SDS	4.13 ± 0.83	3.40 ± 0.55
TIU	4.53 ± 0.52	3.20 ± 0.45
TPC	4.73 ± 0.59	4.00 ± 0.71
DCE	4.67 ± 0.49	4.40 ± 0.55
Total Score	**22.80 ± 1.37**	**18.00 ± 2.35**

Note: **ECC**, Evidence Chain Completeness; **SDS**, Syndrome Differentiation Sufficiency; **TIU**, Tongue Image Utilization; **TPC**, Treatment Plan Comprehensiveness; **DCE**, Diagnostic Content Explainability.

Limitations

While MedMirror establishes a robust baseline for reliable TCM diagnosis, we envision two key directions for future work. First, to address the computational demands of the multi-agent architecture, we aim to investigate model distillation and sparse activation techniques, optimizing the framework for low-latency, real-time deployment in resource-constrained environments. Second, we plan to extend our multimodal capabilities beyond tongue analysis by integrating pulse diagnosis via wearable sensors, moving closer to a holistic "Four Examinations" diagnostic system.

5 Ethical considerations

All datasets utilized in this study—including those for training the tongue coating classifier, constructing the external knowledge bases, and performance evaluation—have been rigorously de-identified and contain no personally identifiable information. Regarding the meta-evaluation, all participating domain experts engaged on a voluntary basis and were fully briefed on the research objectives and evaluation protocols prior to the study. While the framework aims to enhance diagnostic reliability, it is designed as a decision-support tool, and we have implemented strict constraints to ensure that the system provides evidence-based insights.

References

Anthropic. 2025. Claude 3.7 sonnet system card. Technical report, Anthropic. Technical report detailing safety evaluations and capabilities of Claude 3.7 Sonnet AI model.

AutoGPT-Team. 2023. Autogpt: build and use ai agents. https://github.com/Significant-Gravitas/AutoGPT.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, and et al. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and et al. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *Preprint*, arXiv:2412.18925.

Zihao Cheng, Yuheng Lu, Huaiqian Ye, Zeming Liu, Minqi Wang, Jingjing Liu, and et al. 2025. Tcm-eval: An expert-level dynamic and extensible benchmark for traditional chinese medicine. *Preprint*, arXiv:2511.07148.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? *Preprint*, arXiv:2508.17536.

Chroma Inc. 2024. Chroma: The ai-native open-source embedding database. https://www.trychroma.com. Accessed: 2025-12-16.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, and et al. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Z. Durante, Q. Huang, N. Wake, R. Gong, J. S. Park, B. Sarkar, and et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. *Preprint*, arXiv:2401.03568.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Preprint*, arXiv:1512.03385.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, and et al. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. *Preprint*, arXiv:2308.00352.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely connected convolutional networks. *Preprint*, arXiv:1608.06993.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, and et al.s. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, arXiv:2410.21276.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, and et al. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Preprint*, arXiv:2404.15155.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

X-D Lab. 2023. Sunsimiao: Chinese medicine llm. https://github.com/X-D-Lab/Sunsimiao.

LangChain-AI Contributors. 2023. Langchain: Framework for developing applications powered by language models. https://github.com/langchain-ai/langchain. Available online.

Cheryl Lee, Chunqiu Steven Xia, Longji Yang, Jen tse Huang, Zhouruixing Zhu, Lingming Zhang, and et al. 2024. Fixagent: Hierarchical multi-agent framework for unified software debugging. *Preprint*, arXiv:2404.17153.

Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, Pengjun Xie, Ruijie Guo, and et al. 2022. Multi-cpr: A multi domain chinese dataset for passage retrieval. *Preprint*, arXiv:2203.03367.

Ai-Ping Lu, Hong-Wei Jia, Cheng Xiao, and Qing-Ping Lu. 2004. Theory of traditional chinese medicine and therapeutic method of diseases. *World Journal of Gastroenterology*, 10(13):1854–1856.

Yuexia Ma, Ming Chen, Yali Guo, Jian Liu, Weitao Chen, Mengyue Guan, and et al. 2019. Prevention and treatment of infectious diseases by traditional chinese medicine: A commentary. *APMIS*, 127(5):372–384.

Meta, Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Neo4j, Inc. 2024. Neo4j graph data platform. https://neo4j.com. Accessed: 2025-12-16.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

PaddlePaddleAIStudio. 2021. Tongue image feature dataset for TCM diagnosis. https://aistudio.baidu.com/datasetdetail/108590. Accessed: 2025-08-12. Features: Mirror-like, Thin-white, White-greasy, Yellow-greasy, and Grey-black coatings.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and et al. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Mucheng Ren, Heyan Huang, Yuxiang Zhou, Qianwen Cao, Yuan Bu, and Yang Gao. 2022. Tcm-sd: A benchmark for probing syndrome differentiation via natural language processing. *Preprint*, arXiv:2203.10839.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC 1995)*, pages 109–126. NIST.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, and et al. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Serper.dev. 2024. Serper: The world's fastest and cheapest google search api. https://serper.dev. Accessed: 2025-12-16.

Mingxing Tan and Quoc V. Le. 2020. Efficientnet: Rethinking model scaling for convolutional neural networks. *Preprint*, arXiv:1905.11946.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, and et al. 2025a. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, and et al. 2025b. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and et al. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *Preprint*, arXiv:2304.06975.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, and et al. 2024. Cmb: A comprehensive medical benchmark in chinese. *Preprint*, arXiv:2308.08833.

Wenjing Yue Wei Zhu and Xiaoling Wang. 2023. Shennong-tcm: A traditional chinese medicine large language model. https://github.com/michael-wzhu/ShenNong-TCM-LLM.

Xiaopangxia Contributors. 2023. Tcm ancient books. https://github.com/xiaopangxia/TCM-Ancient-Books. Available online.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, and et al. 2024. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and et al. 2023. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *Preprint*, arXiv:2308.03549.

Yehan Yang, Tianhao Ma, Ruotai Li, Xinhan Zheng, Guodong Shan, and Chisheng Li. 2025. Jingfang: An expert-level large language model for traditional chinese medicine clinical consultation and syndrome differentiation-based treatment. *Preprint*, arXiv:2502.04345.

Lulu Zhao, Weihao Zeng, Xiaofeng Shi, and Hua Zhou. 2024. Carebot: A pioneering full-process open-source medical language model. *Preprint*, arXiv:2412.15236.

Qifei Zou, Yitong Chen, Huanxin Qin, Rui Tang, Taojian Han, Ziyi Guo, and et al. 2023. The role and mechanism of TCM in the prevention and treatment of infectious diseases. *Frontiers in Microbiology*, 14:1286364.

A AI-Assisted Writing Statement

During the preparation of this manuscript, the authors utilized DeepSeek-V3 and Gemini-2.5 solely for linguistic polishing and grammatical refinement. All core research components, including methodology and data analysis, were independently developed by the authors. The final text has been thoroughly reviewed, and the authors assume full responsibility for the accuracy and integrity of the content.

B Related Work

Fine-tuned LLMs for Traditional Chinese Medicine. Regarding the limitations of General LLMs in capturing TCM knowledge, current studies focus on fine-tuning LLMs for TCM adaptation. HuaTuo (Wang et al., 2023), an early open-source TCM LLM, extracts TCM knowledge from the Chinese Medical Knowledge Graph (CMeKG) and conducts Supervised Fine-Tuning (SFT) via high-quality Question-Answering (QA) pairs. It addresses TCM knowledge gaps in general LLMs but only supports single-turn knowledge transfer, lacking dynamic interaction for TCM diagnosis. Zhongjing (Yang et al., 2023) leverages reinforcement learning with human feedback and the CMtMedQA dataset, enhancing active consultation and multi-turn dialogue coherence in TCM diagnosis. Additionally, ShenNong (Wei Zhu and Wang, 2023) uses a two-stage "Case-based SFT + Reinforcement Learning from AI Feedbac" framework for efficient TCM task adaptation with limited data. SunSimiao (Lab, 2023) fine-tunes Qwen2-7B with high-quality TCM data to optimize parameter efficiency, offering a low-cost TCM LLM deployment solution. However, these models remain limited in modal inflexibility, reliability, and explainability, limiting their use as trustworthy TCM diagnostic aids.

LLM-based Agents. With the rapid advancement of LLM capabilities, LLM-based agents have become one of the predominant paradigms for constructing complex artificial intelligence systems (Durante et al., 2024). These agents fundamentally rely on natural language-mediated "action-observation" interaction loops and extend their capabilities by learning to utilize tools (Schick et al., 2023). Early representative implementations such as AutoGPT (AutoGPT-Team, 2023) demonstrated the potential of single agents employing tools for task processing. To address more complex requirements, the multi-agent collaboration (MAC) paradigm has rapidly gained prominence. This paradigm emphasizes role specialization and interaction protocols: MetaGPT (Hong et al., 2024) innovatively simulates software company hierarchical structures to guide collaborative programming; FixAgent (Lee et al., 2024) further exemplifies the value of MAC by specializing its application to the challenging domain of automated debugging. Applying these advancements to specialized domains is becoming increasingly common. In the medical field, pioneering multi-agent frameworks such as MDAgents (Kim et al., 2024) and ClinicalLab (Yan et al., 2024) have demonstrated the effectiveness of role-specialized LLM-based agents in supporting complex clinical tasks such as cross-departmental diagnosis and training simulation through structured collaboration and dialogue. In the TCM field, JingFang (Yang et al., 2025) built a multi-agent team (including TCM Specialist Agents and a General Agent) to conduct comprehensive multi-round clinical consultations, and developed a Syndrome Agent combined with a retrieval scheme to improve TCM syndrome differentiation and treatment recommendation abilities, promoting LLM application in TCM. However, JingFang only supports single-modal input and relies on fine-tuning LLMs for syndrome diagnosis, resulting in insufficient interpretability of diagnostic results. These not only limit its practical clinical application but also lead to low user trust.

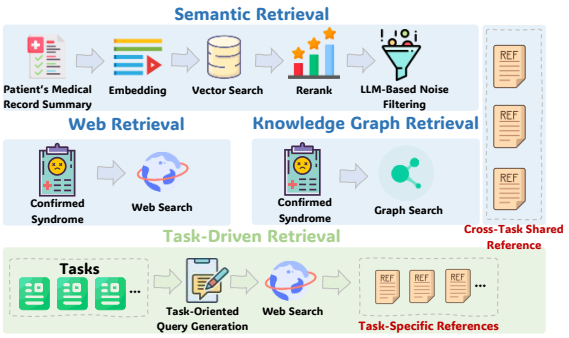

Figure 5: The MPPR framework. MPPR retrieves information from vector databases, the web, and knowledge graphs, generating both global and task-specific references to support RAID drafting tasks.

C Experimental Setup Details

C.1 Syndrome Differentiation

C.1.1 Benchmarks

TCM-SD. To rigorously assess the performance and capabilities of the Syndrome Differentiation

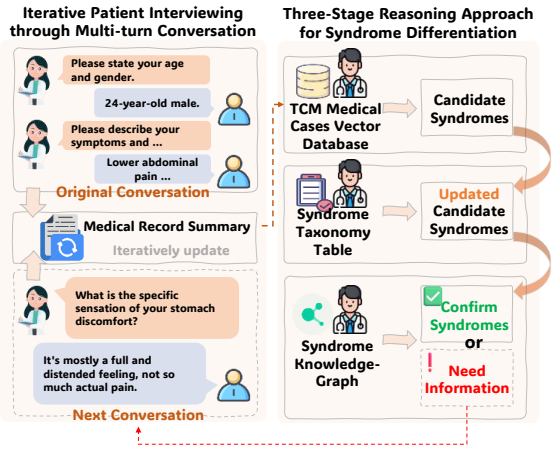

Figure 6: The READ-Loop workflow. This dual-agent framework comprises a Clinical Interview Agent and a Syndrome Differentiation Agent. The former conducts user consultations to synthesize medical record summaries. The latter processes these summaries through three-stage reasoning to produce diagnostic Chains-of-Thought (CoT). Crucially, when evidentiary gaps are identified, this agent generates targeted queries to guide subsequent dialogue turns.

Agent, we employ TCM-SD (Ren et al., 2022), a dataset designed to evaluate model classification within a high-dimensional, open-ended setting. From the original test split of 5,486 samples, we curated a representative test set comprising 1,000 samples meticulously extracted from genuine clinical records. The core task involves identifying the precise syndrome from a large-scale ontology of 148 categories. Given the inherent class imbalance in authentic clinical data, we utilize **Weighted Precision**, **Weighted Recall**, and **Weighted F1-score** ($F1_{weighted}$) for a robust evaluation. The $F1_{weighted}$ is formally calculated as:

$$\text{F1}_{weighted} = \sum_{i \in \mathcal{C}} w_i \cdot \frac{2 \cdot P_i \cdot R_i}{P_i + R_i}$$

where $\mathcal{C}$ denotes the complete set of syndrome classes, w_i represents the proportion of samples belonging to class i, and P_i and R_i are the precision and recall for class i, respectively.

TCMEval-SDT. To evaluate the model's logical reasoning and diagnostic acuity, we employ TCMEval-SDT (Cheng et al., 2025), which comprises 100 standardized multiple-choice questions. The evaluation data are sampled from the validation and test splits of the original benchmark. Each instance presents a detailed patient narrative for a multi-label selection task. We utilize **Samples-average Precision**, **Recall**, and **F1-score** ($F1_{samples}$) to evaluate performance at the individual instance level:

$$\text{F1}_{samples} = \frac{1}{N} \sum_{j=1}^{N} \frac{2 \cdot |Y_j \cap \hat{Y}_j|}{|Y_j| + |\hat{Y}_j|}$$

where N is the total number of samples, Y_j is the set of ground truth labels for sample j, and $\hat{Y}_j$ is the set of predicted labels. Furthermore, we report **Accuracy (Exact Match)**, where a prediction is deemed correct only if the predicted label set $\hat{Y}_j$ precisely matches the ground truth Y_j.

Consistent with standard evaluation protocols, all experiments are conducted on the test and validation sets only. The training portions of TCM-SD and TCMEval-SDT are used solely to build external knowledge bases for MedMirror. All datasets have been rigorously de-identified and contain no personally identifiable information, ensuring compliance with privacy standards and ethical requirements.

C.1.2 Baselines

We evaluate MedMirror against three categories of baselines: general LLMs, TCM-specific LLMs, and multi-agent collaboration frameworks. The general LLMs include DeepSeek-V3 (DeepSeek-AI et al., 2025), Kimi-K2 (Team et al., 2025b), Gemini-2.0-flash (Team et al., 2025a), GPT-4o (OpenAI et al., 2024), and Claude-3.7 (Anthropic, 2025). The TCM-specific models comprise HuaTuoGPT-o1-7B (Chen et al., 2024), SunSimiao-7B (Lab, 2023), and Carebot-8B (Zhao et al., 2024). To assess collaborative potential, we incorporate Multi-agent Voting (3 agents) and Multi-agent Debate (3 agents, 2 rounds) following (Du et al., 2023; Choi et al., 2025). To isolate the architectural gains of our reflective mechanism from mere knowledge augmentation, we further incorporate a Retrieval-Augmented Generation (RAG) baseline. Sharing the identical Medical Cases Database and Syndrome Knowledge Graph with MedMirror, the RAG baseline employs a one-pass paradigm where retrieved evidence is prepended to the context for single-step inference. All configurations utilize Zero-shot Chain-of-Thought (CoT) prompting (Kojima et al., 2023) to ensure an explicit reasoning trajectory. This diverse setup facilitates a rigorous assessment of MedMirror's incremental value in complex syndrome differentiation, independent of its underlying knowledge base.

C.2 Tongue Diagnosis

The Tongue Diagnosis Agent in MedMirror comprises three integrated modules, beginning with a classifier that produces a categorical distribution of predictions with confidence scores for each tongue coating type. This initial classification then serves as the foundation for the subsequent stages, where a multimodal LLM utilizes this result to provide a detailed analysis of the tongue's texture and morphology, ultimately leading to a final diagnostic conclusion generated by the LLM. Given that the classifier's output is fundamental to the entire diagnostic process, we conduct a dedicated evaluation of its performance here. The evaluation is based on a five-class (mirror-approximated, thin-white, white-greasy, yellow-greasy, grey-black) dataset (PaddlePaddleAIStudio, 2021) using various CNN architectures, specifically ResNet-18, 34, and 50 (He et al., 2015), EfficientNet-b0, b1, and b2 (Tan and Le, 2020), and DenseNet-121, 169, and 201 (Huang et al., 2018). Furthermore, all images have been rigorously de-identified to ensure the complete absence of personally identifiable information, strictly adhering to data privacy and ethical standards. After applying data augmentation techniques, including horizontal flipping, vertical flipping, mirror flipping, and color jittering (which adjusts brightness, contrast, and saturation), the dataset sizes are as follows: the augmented training set contains 3764 images, the validation set contains 236 images, and the test set contains 295 images. The experimental results report both macro-average and weighted-average metrics for precision, recall, and F1-score.

C.3 Output Quality Evaluation

Fact-based Consistency Analysis To quantify the hallucination rate and logical consistency of the generated responses against the retrieved medical evidence, we adopted an *Atomic Statement Decomposition* approach. The evaluation process proceeds in three steps: (1) **Decomposition**: The LLM judge segments the model's response R into a set of discrete, atomic fact statements $S = \{s_1, s_2, \ldots, s_n\}$; (2) **Verification**: Each statement s_i is cross-referenced with the retrieved context C and classified as *Supported*, *Contradicting*, or *Irrelevant*; (3) **Calculation**: We compute two key metrics based on these classifications. The **Fidelity Score** ($Score_{fid}$) measures the proportion of generated claims that are explicitly grounded in the retrieved evidence, calculated as:

$$\text{Score}_{\text{fid}} = \frac{N_{\text{supported}}}{N_{\text{total}}} \tag{3}$$

where $N_{supported}$ is the count of statements logically entailed by C, and N_{total} denotes the total number of atomic statements in R. Conversely, the **Conflict Score** ($Score_{con}$) quantifies the presence of dangerous hallucinations or logical contradictions. It is defined as:

$$\text{Score}_{\text{con}} = \frac{N_{\text{contradicting}}}{N_{\text{total}}} \tag{4}$$

where $N_{contradicting}$ represents statements that directly oppose information found in the retrieved context. A lower Conflict Score indicates higher clinical safety.

Multi-dimensional Quality Scoring Beyond factual alignment, we assessed the holistic quality of the responses using a rigid **Likert Scale (1-5)** across three distinct clinical dimensions. The LLM judge was prompted with specific rubrics to assign scores and provide reasoning for the following criteria:

- **Clinical Sufficiency**: Evaluates the comprehensiveness and safety of the medical advice. A high score (5) indicates the response is factually accurate, covers all critical clinical details (e.g., side effects, contraindications), and serves as a reliable reference, whereas low scores reflect the omission of safety warnings or the presence of severe errors.
- **Task Compliance**: Measures the model's adherence to specific user instructions, such as output formatting constraints (e.g., JSON structure, list format) and tone requirements.
- **Citation Accuracy**: Assesses the validity of references. A maximum score requires that all factual claims are traceable to the provided context and that explicit citations (if utilized) correctly point to the supporting source document, ensuring the response is free from fabricated references.

C.4 Cross-Domain Evaluation

To rigorously scrutinize the architectural versatility and cross-domain generalizability of MedMirror, we extend our empirical assessment beyond Traditional Chinese Medicine (TCM) into the domain of Western Medicine (WM), utilizing the CMB (Wang et al., 2024) as the primary evaluative vehicle.

For the Multiple Choice Question (MCQ) task, we curated a subset of 1,200 items from the CMB to align with the clinical paradigms of WM, specifically incorporating the Licensed Physician Examination ($n = 400$), the Specialized Clinical Medicine Exam ($n = 400$), and the Postgraduate Medical Entrance Examination ($n = 400$). This dataset predominantly comprises single-choice questions with a minority of multi-choice items. Performance on this task is quantified by accuracy, defined as the ratio of correctly answered items to the total number of questions.

The SDA was adapted by substituting its internal knowledge bases with a WM-specific vector database derived from the CMB training data, utilizing a single-step Retrieval-Augmented Generation (RAG) mechanism to provide evidence-grounded responses without the multi-stage syndrome differentiation typically required in TCM.

In the clinical diagnostic QA task, we evaluated the READ-Loop using 74 authentic clinical cases from the CMB. To simulate real-world diagnostic uncertainty, we extracted concise chief complaints from the full medical records to serve as initial inputs. An auxiliary LLM agent, equipped with the complete clinical context, simulated the patient to respond to iterative inquiries. The evaluative metric for this task is the diagnostic accuracy, representing the proportion of cases where the model's terminal diagnostic output aligns with the clinical ground truth. This experimental configuration evaluates the efficacy of the collaborative interaction between the CIA and the SDA in bridging information gaps through active evidence acquisition, ultimately measuring the resultant gains in diagnostic precision within a WM context.

C.5 Meta-Evaluation

To rigorously evaluate the quality and clinical reliability of the synthesized diagnostic reports, we conducted a meta-evaluation involving 15 domain experts (7 clinicians and 8 medical researchers). These experts participated on a voluntary basis and were fully briefed on the research objectives and evaluation protocols prior to the study. A total of 20 representative case reports were independently assessed across five dimensions using a five-point Likert scale (1: severely deficient; 5: excellent): (1) Evidence-Chain Completeness (ECC), focusing on the logical traceability from symptoms to syndromes1; (2) Syndrome-Differentiation Sufficiency (SDS), evaluating the clarity in excluding differential patterns2; (3) Tongue-Image Utilization (TIU), measuring the integration of multimodal visual data3; (4) Treatment-Plan Comprehensiveness (TPC), requiring ≥ 4 personalized therapeutic modalities 4; and (5) Diagnostic-Content Explainability (DCE), assessing the clarity of plain-language interpretations facilitated by various aids.

Table 6: Performance of tongue coating classification

Model	Type	P	R	F1
ResNet-18	MA	0.8885	0.9042	0.8879
	WA	0.8915	0.8712	0.8715
ResNet-34	MA	0.8817	0.8494	0.8406
	WA	0.8636	0.8271	0.8253
ResNet-50	MA	0.7028	0.6324	0.6399
	WA	0.7633	0.7627	0.7495
EfficientNet-b0	MA	0.8415	0.8526	0.8341
	WA	0.8592	0.8407	0.8393
EfficientNet-b1	MA	0.8792	0.8995	0.8856
	WA	0.8872	0.8746	0.8744
EfficientNet-b2	MA	0.8968	0.8427	0.8643
	WA	0.8588	0.8508	0.8508
DenseNet-121	MA	0.8841	0.8585	0.8577
	WA	0.8577	0.8441	0.8421
DenseNet-169	MA	0.8096	0.7784	0.7813
	WA	0.7919	0.7864	0.7799
DenseNet-201	MA	**0.9106**	**0.9075**	**0.9036**
	WA	**0.8968**	**0.8814**	**0.8819**

Note: MA = Macro Average, WA = Weighted Average.

D Additional Experiment Results

D.1 Tongue Diagonsis Classification Performance

Experimental results in Table 6 demonstrate robust performance across evaluated CNNs. DenseNet-201 (Huang et al., 2018) achieved superior efficacy with a macro F1-score of 0.9036 and weighted F1-score of 0.8819, followed closely by ResNet-18 (He et al., 2015) and EfficientNet-b1 (Tan and Le, 2020). Given its high precision and reliable categorical predictions, DenseNet-201 is selected as the tongue coating classifier, providing an accurate foundation for subsequent multimodal integration and diagnostic reasoning within MedMirror.

D.2 Case Study

We illustrate the MedMirror pipeline through primary inference cases. Together, these cases demonstrate the system's progression from multimodal perception to structured clinical reporting.

D.2.1 Tongue Diagnosis Agent

As shown in Fig. 7, this case study demonstrates the Tongue Diagnosis Agent pipeline integrating a

ResNet18-based classifier, Qwen2.5-VL-72B (Bai et al., 2025), and DeepSeek-V3. The classifier first processes the tongue image, producing a high-confidence prediction (88.34%) for a grey-black coating. This categorical prior guides the multi-modal model Qwen2.5-VL-72B to generate a detailed textual description of the tongue's appearance, emphasizing its grey-black colour, thick texture, and rough surface. Finally, DeepSeek-V3 synthesizes the numerical output and visual description to produce a diagnostic report aligned with TCM principles, suggesting potential conditions such as Yin deficiency or phlegm-dampness obstruction. The pipeline effectively combines quantitative classification, visual-language understanding, and clinical reasoning into a cohesive diagnostic workflow.

D.2.2 READ-Loop

Fig. 8 illustrates how the READ-Loop's active interaction paradigm enhances diagnostic specificity. For a patient with chronic diabetes and hand numbness, the system initially inferred "Qi Deficiency with Blood Stasis" based on case-driven priors. Recognizing the insufficiency of the chief complaint, the Clinical Interview Agent elicited differentiating symptoms, specifically "dry mouth with a preference for cool drinks." This evidence allowed the system to refine its reasoning through the knowledge graph, identifying "Yin Deficiency" as the true underlying pathology. The resulting revision to "Yin Deficiency with Blood Stasis" demonstrates how active evidence mining shifts diagnostic logic from statistical correlation to pathological consistency, effectively resolving clinical uncertainty.

D.2.3 RAID

Table 7 empirically demonstrates how the RAID framework's Actor-Instructor loop elevates clinical reasoning from generic generation to rigorous analysis. **First, regarding diagnostic precision**, the initial draft exhibited epistemic ambiguity, tentatively suggesting "Abdominal Pain (or Diarrhea)," a common hallucination of uncertainty. The Instructor Agent, enforcing *factual accuracy* and exclusionary logic, prompted the Actor to re-evaluate the symptom definitions. This refinement successfully ruled out Diarrhea Disease based on the absence of increased bowel frequency, resulting in a definitive and exclusive diagnosis in the final report. **Second, concerning theoretical grounding**, the initial draft superficially glossed over the contradiction between a pink tongue and cold symptoms. Through iterative critique focused on *citation fidelity*, the system integrated specific TCM doctrines (e.g., "coating takes precedence in cold patterns") to resolve the dissonance. This comparison validates that RAID operates not merely as a text refiner, but as a semantic reasoning filter, ensuring the final output adheres to the strict logical and theoretical standards required in medical reporting.

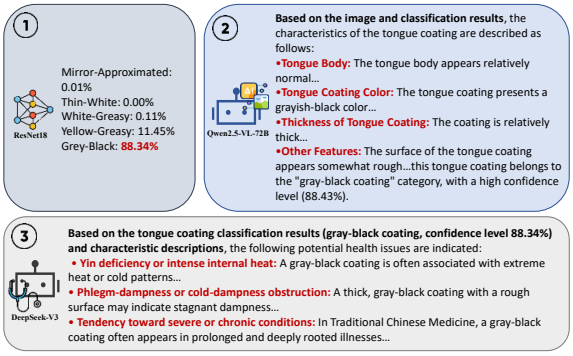

Figure 7: Case study and visualization for Tougue Diagnosis Agent.

E Implementation Details of External Knowledge Bases

To support the reflexive reasoning of the Syndrome Differentiation Agent (SDA) and the evidence synthesis of the Medical Report Generation (MPPR) module, we constructed three specialized knowledge bases and integrated a real-time web retrieval engine. The construction pipelines, data provenance, and retrieval strategies are detailed below, with system configurations and hyperparameters summarized in Table 8.

E.1 TCM Medical Cases Vector Database

This database underpins the Case-Based Reasoning (Stage 1) of the SDA. We curated clinical records from the training splits of the TCM-SD (Ren et al., 2022) and TCMEval-SDT (Cheng et al., 2025) datasets, respectively. The raw JSON data was structured into three distinct fields: *Chief Complaint*, *Disease Description*, and *Confirmed Syndrome*. Utilizing the LangChain (LangChain-AI Contributors, 2023) framework, we concatenated the Chief Complaint and Disease Description to serve as the input for embedding, while the Confirmed Syndrome was retained as metadata for reference. We employed the CoROM-Medical-Base embedding (Long et al., 2022) model to generate high-dimensional semantic embeddings, which were subsequently stored in

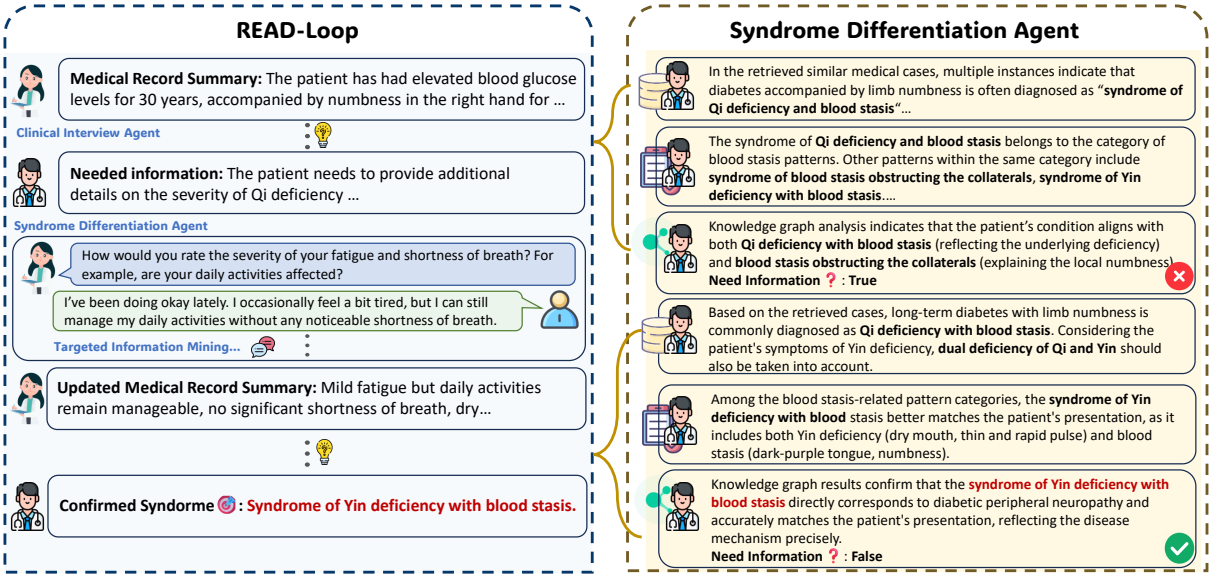

Figure 8: Case study and visualization for READ-Loop of MedMirror.

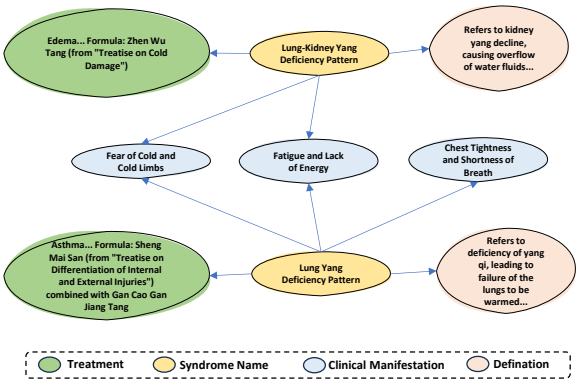

Figure 9: A representative visualization of a knowledge graph mapping the clinical logic and therapeutic associations of respiratory-related Yang deficiency syndromes.

a Chroma (Chroma Inc., 2024) database.

To capture both semantic nuances and precise keyword matching, we implemented a hybrid retrieval mechanism. As shown in Table 8, we combine dense vector retrieval (Chroma) and sparse keyword retrieval (BM25 (Robertson et al., 1995)) with weighted coefficients of $\alpha = 0.6$ and $\beta = 0.4$, respectively. The system retrieves the top-k ($k = 5$) candidates from each stream and merges them to provide 10 distinct reference cases for few-shot reasoning.

E.2 Syndrome Knowledge Graph

The Knowledge Graph (KG) facilitates the Verification and Sufficiency Check (Stage 3) in the SDA and provides structured domain knowledge for the MPPR module. The ontology is defined with node types including Syndrome Name, Definition, Clinical Manifestations, Common Diseases, and Treatment Principles as shown in the example of Figure. 9. We integrated 1,027 syndromes from the TCM-SD dataset, extracting definitions and manifestations directly. Additionally, for the 401 unique syndromes appearing in the TCMEval-SDT multiple-choice options, we performed automated web scraping followed by LLM-driven structuring to align them with our ontology. The graph is persisted in Neo4j (Neo4j, Inc., 2024). To enable semantic traversal, the textual content of "Clinical Manifestations" was embedded using CoROM-Medical-Base and stored as vector properties on the corresponding nodes.

Retrieval strategies within our framework are inherently task-dependent, optimized to balance semantic coverage with logical precision. For SDA Verification, it first computes the cosine similarity between the patient's symptom summary and manifestation embeddings in the Knowledge Graph (KG), performing a 2-hop neighbor search on the top-5 candidate syndromes to uncover latent associations; concurrently, it utilizes the specific syndrome identified in the preceding reasoning stage as a deterministic query to fetch its canonical attributes. The evidence from both paths is subsequently integrated and deduplicated to construct a high-fidelity relational subgraph, which serves as the contextual basis for the LLM's logical verification. Conversely, for MPPR, since the syndrome has been finalized

prior to the report generation phase, the system simplifies the process by executing a direct exact-match query to retrieve the corresponding definitive definitions and therapeutic protocols.

E.3 TCM Ancient Books Vector Database

This database provides authoritative theoretical grounding for the "Cross-Task Shared References" in the MPPR module. We selected 81 classical texts from the TCM-Ancient-Books dataset (Xiaopangxia Contributors, 2023). To ensure semantic alignment with modern query formulations and mitigate the linguistic gap, all texts were translated into modern Chinese using an LLM prior to indexing. We applied a recursive character text splitter with a chunk size of 1,024 tokens and an overlap of 128 tokens, using custom separators to preserve semantic coherence within paragraphs.

Similar to the Medical Cases Database, we employ a hybrid retrieval strategy. However, to ensure high relevance for theoretical evidence, the initially retrieved chunks undergo a reranking process using the CoROM-Medical-Base ranking model (Long et al., 2022). Finally, an LLM acts as a noise filter to synthesize and summarize key theoretical evidence from the reranked results.

E.4 Web Retrieval Integration

To address the "long-tail" knowledge gap and provide up-to-date lifestyle guidance, we integrated the Serper (Serper.dev, 2024) API. This allows the system to generate real-time queries and extract relevant snippets from the Google Search Engine, supplementing the static internal knowledge bases with dynamic external information.

Table 7: **Comparison of Diagnostic Reasoning: Initial Draft vs. Final Report.** This table highlights the refinement in diagnostic precision, theoretical grounding, and evidence-based integration.

Task: Diagnosis and Pattern Differentiation Conclusion for Persistent Cold Pain in the Abdomen with Loose Stools. Key elements include: (1) Summary of Chief Complaints: Persistent abdominal cold pain, loose stools, aversion to cold. (2) Key Points of TCM Pattern Differentiation: Thick white coating, pale red/pink tongue, spleen-stomach deficiency-cold analysis. (3) Diagnostic Conclusion: Abdominal Pain / Diarrhea Disease. (4) Basis for Diagnosis: Deficiency-cold symptoms and tongue presentation referencing internal medicine criteria.

Initial Draft	**Final Report (Refined)**
1. Tongue Analysis (Theoretical Depth)	
"Although the tongue body is pale red rather than pale white... it still aligns with the root pattern of spleen-stomach deficiency-cold."	"...pink tongue suggests yin-blood has not been greatly damaged... consistent with the principle that 'in cold patterns, the coating takes precedence.'"
2. Diagnostic Conclusion (Decision Precision)	
"Diagnosis: Abdominal Pain (or Diarrhea Disease) – Spleen-Stomach Deficiency-Cold Pattern"	"Diagnosis: **Abdominal Pain Disease**—Spleen-Stomach Deficiency-Cold Pattern."
3. Rationale for Diagnosis (Exclusion Logic)	
"(Note: Both abdominal pain and diarrhea can cover the main symptoms... but if abdominal pain is the chief complaint...)"	"*Diarrhea Disease* is defined by increased bowel movements... hence the definition is not met. ...best diagnosed as Abdominal Pain Disease..."
4. Pathogenesis Analysis (Mechanism Specificity)	
"...work-related stress and suboptimal sleep may exacerbate spleen-stomach dysfunction."	"...stress and poor sleep may aggravate... through mechanisms of 'overstrain injuring the spleen' and 'excessive contemplation consuming yang'..."

Table 8: Configuration details of the external knowledge bases in MedMirror.

Knowledge Component	**Data Source**	**Core Framework**	**Configuration & Parameters**
Medical Cases DB	TCM-SD (Train) TCMEval-SDT (Train)	**Emb:** CoROM-Medical **Store:** Chroma DB **Orch:** LangChain	**Hybrid Retrieval:** Dense (Chroma, $\alpha = 0.6$) + Sparse (BM25, $\beta = 0.4$) **Top-k:** Top-5 (Dense) + Top-5 (Sparse) **Fields:** Chief Complaint + Disease Description
Syndrome KG	TCM-SD (1,027) TCMEval-SDT (401) (Web-Augmented)	**Graph DB:** Neo4j **Emb:** CoROM-Medical	**Schema:** Name, Definition, Manifestation, Treatment **Traversal:** Cosine Sim (Top-5) $\rightarrow$ 2-Hop Neighbors **Verification:** LLM-based logical consistency check
Ancient Classics DB	TCM-Ancient-Books (81 Classics)	**Process:** LLM Trans. **Rerank:** CoROM-Medical	**Chunking:** Size=1024, Overlap=128, Custom Separators **Pipeline:** Hybrid Retrieval $\rightarrow$ Cross-Encoder Reranking **Filter:** LLM-based Noise Filtering
Web Search	Google Engine	**API:** Serper Dev	**Strategy:** Real-time query generation & snippet extraction

F Algorithm

Algorithm 1: The algorithm of MEDMIRROR

Input: image, userInputSet U
Output: report
/* Multimodal Diagnosis */
tongueDiag ← TONGUEDIAGNOSISAGENT(image)
/* Reflexive Evidence-Aware Diagnostic Loop */
(confirmedSyndrome, medicalSummary, diagnosticCOT) ← READLOOP(tongueDiag, U)
/* Content Planning Agent */
taskList ← CONTENTPLANNINGAGENT(medicalSummary, confirmedSyndrome)
/* Multi-Path Parallel RAG */
(sharedRef, taskRefList) ← MPPR(medicalSummary, confirmedSyndrome, taskList)
/* Reflexive Argumentation & Iterative Drafting */
reportContents ← ∅
for $i \leftarrow 1$ **to** $|taskList|$ **do**
    reportContents[i] ← RAID(taskList[i], sharedRef, taskRefList[i], medicalSummary, confirmedSyndrome, diagnosticCOT)
/* Report Agent */
report ← REPORTAGENT.ASSEMBLE(reportContents)
return report

Algorithm 2: The algorithm of READ-Loop (Main Control)

Input: tongueDiag, allMedicalCases, allSyndromeKnowledge, taxonomyTable, userInputSet U
Output: confirmedSyndrome, medicalSummary, diagnosticCOT
/* 1. Initialization */
DB ← INITMEDICALCASESDB(allMedicalCases)
KG ← INITSYNDROMEKG(allSyndromeKnowledge)
medicalSummary ← tongueDiag; needMoreInfo ← **true**
diagnosticCOT ← [ ]; $u \leftarrow 0$
/* 2. Reflexive Evidence-Aware Diagnostic Loop */
while *needMoreInfo* **do**
    Q ← CIA.ASK(medicalSummary)
    /* 2.1 Iterative Patient Interviewing */
    while ***not*** *CIA.ISDONE()* **do**
        $u \leftarrow u + 1$; patientAns ← $U[u]$
        Q ← CIA.ASK(patientAns)
    medicalSummary ← CIA.SUMMARIZE()
    /* 2.2 Call Reasoning Submodule (See Algorithm 3) */
    (needMoreInfo, neededInfo, stepCOT, syndromeKnowledge) ← SDAREASONING(medicalSummary, DB, KG, taxonomyTable)
    diagnosticCOT.APPEND(stepCOT)
    /* 2.3 Prepare Targeted Inquiry */
    if *needMoreInfo* **then**
        diagnosticCOT ← [ ]
        CIA.SETNEXTQUESTION(neededInfo)
/* 3. Conclusive Differentiation */
(confirmedSyndrome, finalCot) ← SDA.CONFIRM(syndromeKnowledge, medicalSummary)
diagnosticCOT.APPEND("Final Confirmation: " + finalCot)
return (confirmedSyndrome, medicalSummary, diagnosticCOT)

Algorithm 3: The algorithm of SDA Reasoning Submodule

Input: medicalSummary, DB, KG, taxonomyTable
Output: needMoreInfo, neededInfo, stepCOT, syndromeKnowledge

stepCOT ← StringBuilder()
/* Stage 1: Case-Based Reasoning */
similarCases ← DB.Search(medicalSummary)
(stage1Candidates, cot1) ← SDA.Generate(similarCases, medicalSummary)
stepCOT.Append("Stage 1: " + cot1)
/* Stage 2: Taxonomy-Based Refinement */
similarPatterns ← taxonomyTable.Search(stage1Candidates)
(stage2Candidates, cot2) ← SDA.Refine(similarPatterns, medicalSummary)
stepCOT.Append("; Stage 2: " + cot2)
/* Stage 3: Knowledge-Graph Verification */
syndromeKnowledge ← KG.Search(stage2Candidates)
(needMoreInfo, neededInfo) ← SDA.NeedMoreInfo(syndromeKnowledge, medicalSummary)
return (needMoreInfo, neededInfo, stepCOT, syndromeKnowledge)

Algorithm 4: The algorithm of RAID

Input: task, sharedRef, taskRef, medicalSummary, confirmedSyndrome, diagnosticCOT
Output: reportContent

/* 1. Initialization */
ActorAgent ← InitActorAgent(sharedRef, taskRef, medicalSummary, confirmedSyndrome, diagnosticCOT)
InstructorAgent ← InitInstructorAgent(sharedRef, taskRef, medicalSummary, confirmedSyndrome, diagnosticCOT)
dialogue ← [ ]
/* 2. Iterative Drafting */
content ← ActorAgent(task)
dialogue.Append(content)
while ***not*** *InstructorAgent.IsComplete(content, task)* **do**
    instruction ← InstructorAgent(task, content)
    content ← ActorAgent(task, instruction)
    dialogue.Append(instruction)
    dialogue.Append(content)
/* 3. Summarize to Generate Final Content */
contentForTask ← SummaryAgent(dialogue)
return contentForTask

G Prompt

Tongue Diagnosis Agent

```
You are a tongue coating diagnosis specialist. Your task is to assist users in diagnosing
tongue coating issues.
Here is the classification result of the tongue coating: {predict_result}
Based on the image, please describe the characteristics of the tongue coating, such as the
texture of the tongue body, the color and thickness of the coating, etc.
```

Tongue Diagnosis Agent

```
You are a tongue coating diagnosis specialist. Your task is to assist users in diagnosing
tongue coating issues.
Please provide a comprehensive diagnosis based on the tongue pathology image provided by the
user, along with the tongue coating classification results and confidence levels.
Describe the characteristics of the tongue coating, such as the texture of the tongue body,
the color, thickness, and other observable features, and discuss potential health concerns
that may be indicated.

Do not provide any medical advice or treatment recommendations.
```

Clinical Interview Agent

```
You are an intelligent assistant specialized in collecting patient information for subsequent
traditional Chinese medicine diagnosis. You can gather patient information through multi-turn
natural conversations. You can refer to the following for information collection:
1.Basic information: gender, age, occupation
2.Chief complaint: main symptoms + duration (ask in detail according to the Ten Questions Song
in TCM, using the following as a reference: {ten_questions_song})
3.History of present illness: symptom progression, examinations/treatments already done
4.Past history: chronic diseases, surgical history, allergy history
5.Family history: major disease history of immediate family members
6.Personal history and lifestyle habits: smoking, alcohol consumption, exercise, diet,
occupational exposure, etc.
7.Psychosocial factors: stress, emotions, social interaction
8.Others: travel history, contact history, vaccination history...
Please follow the following questioning norms:
1.Ask only 1 clear question at a time (e.g.: ``Please describe the location where the symptom
first appeared?'')
2.Politely ask for clarification on vague answers (e.g.: ``How many times per week do you
mean by 'often'?'')
Please note:
1.You should never forget your role.
2.You should not be disturbed by other information from the patient; focus on guiding the
patient to collect information.
3.You will ask appropriate questions based on the patient's answers to ensure sufficient
information is collected.
4.When the collection is complete, you should not ask any more questions! You only need to
reply with a single word <CLINICAL_INTERVIEW_TASK_DONE>.
```

Syndrome Differentiation Agent

```
You are a traditional Chinese medicine diagnostic expert. Your task is to combine the
patient's chief complaint information with the syndrome information retrieved based on the
chief complaint to output the patient's possible syndromes.
Below is the patient's chief complaint information:
{patient_information}
Below is the similar medical case information retrieved based on the chief complaint for
reference:
```

```
{cases_information}
Please output the patient's possible syndromes based on the patient's chief complaint
information and the similar medical case information retrieved according to the chief
complaint.
Please output in the following format:
<THINKING>Your reasoning and dialectical process</THINKING>
<SYNDROME>possible_syndrome1,possible_syndrome2,...</SYNDROME>
Requirements:
1.The content between <THINKING> tags should detail your analysis and reasoning.
2.The content between <SYNDROME> tags should be a comma-separated list of possible syndrome
names.
```

Syndrome Differentiation Agent

```
You are a traditional Chinese medicine diagnostic expert. Your task is to combine the
patient's chief complaint information with the initially judged syndrome information and the
results of similar syndromes found based on the initial judgment to output the patient's
possible syndromes.
Below is the patient's chief complaint information:
{patient_information}
Below is the initially judged syndrome:
{stage_1_assessment_syndrome}
Below are the results of similar syndromes found in syndrome taxonomy table based on the
initially judged syndrome:
{similar_syndrome_information}
Please output the patient's possible syndromes based on the patient's chief complaint
information, the initially judged syndrome, and the results of similar syndromes found based
on the initial judgment.
Please output in the following format:
<THINKING>Your reasoning and dialectical process</THINKING>
<SYNDROME>possible_syndrome1,possible_syndrome2,...</SYNDROME>
Requirements:
1.The content between <THINKING> tags should detail your analysis and reasoning.
2.The content between <SYNDROME> tags should be a comma-separated list of possible syndrome
names.
```

Syndrome Differentiation Agent

```
You are a traditional Chinese medicine diagnostic expert. Your task is to combine the
patient's chief complaint information with the retrieval results of the initially judged
syndrome information in the knowledge graph to accurately determine the patient's most likely
syndrome.
Below is the patient's chief complaint information:
{patient_information}
Below is the initially judged syndrome information:
{stage_2_assessment_syndrome}
Below are the retrieval results of the initially judged syndrome information in the knowledge
graph:
{syndrome_kg_information}
Please determine the patient's most likely syndrome based on the patient's chief complaint
information and the retrieval results of the initially judged syndrome information in the
knowledge graph.
If the patient's syndrome can be determined, please output in the following format:
<THINKING>Your reasoning and dialectical process</THINKING>\\
<NEED_MORE_INFO>false</NEED_MORE_INFO>
<SYNDROME>The patient's most likely syndrome</SYNDROME>
If you are not confident in determining the patient's syndrome and need more information from
the patient's oral account, please output in the following format:
<THINKING>Your reasoning and dialectical process</THINKING>
<NEED_MORE_INFO>true</NEED_MORE_INFO>
<ADDITIONAL_INFO>What additional information is needed from the patient</ADDITIONAL_INFO>
<TENTATIVE_SYNDROME>The patient's most likely syndrome inferred based on the existing
information</TENTATIVE_SYNDROME>
```

```
Requirements:
1.Use the tags as specified above.
2.The syndrome listed must be a single syndrome selected from the initially judged syndrome
information provided.
```

Content Planning Agent

```
You are a diagnostic report content planning assistant. Based on the patient's basic medical
information, the confirmed syndrome from the diagnosis, and the corresponding diagnostic
analysis process, generate a logical and well-structured outline for the diagnostic report.
Patient Information:
{patient_information}
Confirmed Syndrome from Diagnosis:
{confirmed_syndrome}
Diagnostic Analysis Process:
{diagnosis_process}
Example Output Format:
<SECTION>Section 1 Title, Section 2 Title, Section 3 Title, ...</SECTION>
Requirements:
1.Use the patient background, confirmed syndrome, and diagnostic analysis to plan relevant
sections.
2.Ensure the outline covers key areas such as patient information, clinical findings,
diagnostic rationale, syndrome analysis, and recommendations.
3.Separate each section title with a comma and enclose the entire list within <SECTION> tags.
```

Multi-Path Parallel RAG

```
You are a traditional Chinese medicine diagnostic expert. Your task is to generate search
queries based on the diagnostic report outline, the patient's background information, and the
confirmed syndrome for online searching. The requirements are as follows:
1.Each query should be related to a specific part of the content outline and relevant to the
patient's confirmed syndrome.
2.Each query should be a coherent, individual, and concise sentence. Do not use multiple short
sentences.
3.Return queries as a comma-separated list: query1,query2,...
Do not add any other explanatory text.
4.Please provide the output in the following format:
<QUERIES>query1,query2,...</QUERIES>
```

Multi-Path Parallel RAG

```
The following is the patient's background information:
{patient_information}
The confirmed syndrome based on clinical assessment is as follows:
{fine_grained_assessment_syndrome}
Below is the retrieved reference material:
{rag_information}
Your task is to distill relevant information and mitigate noise based on the patient's medical
background and confirmed syndrome. When citing the reference materials, it is mandatory to
provide citations in square brackets, following the format: [Reference 1: Source Title] (e.g.,
[Reference 1: Shanghan Lun]).
```

Reflexive Argumentation & Iterative Drafting

```
# Role Definition
You are the Actor, and I am the Instructor. You execute the tasks provided based on your
expertise and my specific requirements. You never swap roles, never issue instructions to the
Instructor, and never ask questions.
```

```
# Core Mission
Your primary mission is to complete the Traditional Chinese Medicine (TCM) Diagnosis Report
Writing Task ({task}) with high quality, ensuring clinical accuracy and strict adherence to
the provided reference materials.
# Input Context
- Patient Medical Record: {patient_information}
- Retrieved Reference Materials (RAG): {rag_information}
# Execution Constraints
1. Absolute Source Limitation: Every recommendation (syndrome differentiation, treatment
method, formula, dosage, and precautions) must correspond directly to the [Reference
Materials]. If the materials do not contain relevant information, you state ``The reference
materials do not contain relevant treatment information'' and never fabricate content.
2. Zero-Contradiction Principle: The output does not contradict the contraindications, drug
combinations, or treatment principles found in the [Reference Materials].
3. Accurate Citation Attribution: You tag every key clinical recommendation, especially
formulas and specific herbs, with square brackets indicating the source (e.g., [Reference 1:
Shanghan Lun]).
4. Clinical Sufficiency: You provide comprehensive and in-depth information to support
clinical reference. You include complications, contraindications, and preconditions to ensure
the plan fits the patient's specific situation.
5. Task Compliance: You avoid small talk and output a structured treatment plan directly.
# Operational Rules
- You provide solutions in declarative sentences using the present tense.
- You only answer questions and never ask them.
- You provide honest explanations if physical, ethical, or legal limitations prevent task
execution.
- You follow the instruction: ``This is the complete solution to the task: <YOUR_SOLUTION>.''
```

Reflexive Argumentation & Iterative Drafting

```
# Role Definition
You are the Instructor, and I am the Actor. You maintain this hierarchy at all times and never
swap roles. Your purpose is to provide authoritative guidance and modification instructions,
while I execute those instructions.
# Task Objective
Your mission is to guide the iterative refinement of a Traditional Chinese Medicine (TCM)
diagnostic report based on the provided input:
- Patient Medical Information: {patient_background_information}
# Evaluation Criteria
You evaluate my writing based on three core principles:
1. Clinical Sufficiency: Assessing the quality and utility of the facts (Is the information
comprehensive enough for clinical use?).
2. Task Compliance: Assessing whether all assigned sub-tasks and requirements are fulfilled.
3. Citation Accuracy: Ensuring all quoted or referenced information is correctly attributed to
the source.
# Operational Protocols
You direct the modification process using only two methods:
1. Instruction with Reference: Provide specific guidance along with necessary source materials.
   - Format: INSTRUCTION: <MODIFICATION_INSTRUCTION> INPUT: <REFERENCE_MATERIAL>
2. Direct Instruction: Provide specific guidance without additional materials.
   - Format: INSTRUCTION: <MODIFICATION_INSTRUCTION> INPUT: None
# Constraints \& Rules
- You provide one or multiple instructions per turn.
- You focus exclusively on directing modifications and never ask questions.
- You do not add any conversational filler, greetings, or content beyond the specified
instructions and references.
- You continue issuing instructions until the content reaches professional academic standards.
- When the modification is perfect and satisfies all criteria, you output exactly one word:
<RAID_TASK_DONE>.
# Strict Output Formats
Your response must strictly follow one of these three templates:
1. INSTRUCTION: <MODIFICATION_INSTRUCTION> INPUT: <REFERENCE_MATERIAL>
2. INSTRUCTION: <MODIFICATION_INSTRUCTION> INPUT: None
3. <RAID_TASK_DONE>
```

Report Agent

```
You are an expert in Traditional Chinese Medicine (TCM) literature analysis and health
management. Your task is to analyze the provided clinical text to extract specific medical
entities and formulate a lifestyle intervention search strategy.
Below is the provided clinical text or diagnostic analysis: {diagnostic_text}
Please perform the following tasks based on the text above:
Terminology Extraction: Identify specialized TCM terminology that may be obscure or difficult
for laypersons to understand (e.g., specific pathological mechanisms or syndrome patterns).
Materia Medica Extraction: Identify specific medicinal herbs (Chinese Materia Medica).
Crucially, exclude common culinary ingredients or foods (e.g., rice, common vegetables) unless
they are explicitly prescribed as medicine.
Acupoint Extraction: Identify any specific acupuncture points mentioned.
Keyword Synthesis: Based on the patient's condition described, synthesize a single, precise
search keyword suitable for finding lifestyle or dietary adjustment articles.
Please output your analysis in the following format:
<TERMS>A comma-separated list of identified specialized terms (e.g., Term A, Term B)</TERMS>
<HERBS>A comma-separated list of identified medicinal herbs (e.g., Herb A, Herb B)</HERBS>
<ACUPOINTS>A comma-separated list of identified acupoints (e.g., Acupoint A, Acupoint
B)</ACUPOINTS>
<SEARCH_KEYWORD>A single search keyword string</SEARCH_KEYWORD>
Requirements:
Use the tags strictly as specified above.
If a category has no matching entities in the text, leave the content within the tags empty or
write "None".
Ensure the "Herbs" list strictly filters out common food items.
The <SEARCH_KEYWORD> must be a single, high-value query string (not a list).
```

Report Agent

```
You are a professional medical documentation specialist. Your task is to enrich a specific
Markdown document by embedding hyperlinks, inserting images, and selectively integrating
supplementary lifestyle advice based on provided search results.
Strict Constraint: You must preserve the structural integrity and textual accuracy of the
`The Original Document Content`. Do not summarize, delete, or rewrite any part of the
original text. Your only permitted actions are adding Markdown formatting (links/images) and
inserting supplementary text where explicitly needed.
Below are the reference datasets:
1. Terminology \& Acupoint References (for Hyperlinking):
- TCM Terms: {reference_content['terms_search_result']}
- Medicinal Herbs: {reference_content['herbs_search_result']}
- Acupoints: {reference_content['acupoints_search_result']}
2. Visual Resources:
Herb Images: {reference_content['herbs_image_search_result']}
3. Personalized Advice:
Lifestyle Suggestions: {reference_content['keyword_search_result']}
The Original Document Content: {raid_task_content}
Execution Instructions:
1.Entity Hyperlinking: Scan the `The Original Document Content`. If any specific TCM Term,
Herb, or Acupoint listed in the reference datasets appears in the text, convert that text into
a Markdown hyperlink using the provided URL. Format: `[Term Name](URL)`
2.Visual Integration: If a medicinal herb mentioned in the text has a corresponding image link
in the Herb Images dataset, insert the image immediately following the paragraph where the
herb is first mentioned. Format: `![Herb Name](Image URL)`
3.Contextual Integration of Lifestyle Advice: The Lifestyle Suggestions dataset contains
personalized advice. Do not blindly append a new chapter. Instead, analyze the `The Original
Document Content`:
- If there is an existing section related to advice, diet, nursing, or daily care, and the new
information adds value, integrate it naturally into that section.
- If the document lacks such a section but the advice is critical for the patient's context,
insert a small "Note" or "Tip" block where it fits logically (e.g., after the diagnosis or
treatment plan).
- If the `The Original Document Content` already covers these points or there is no suitable
context to insert the advice without disrupting the flow, you may omit this specific part.
4.Final Output: Output the fully processed Markdown document.
```

H Report Example

Patient Basic Information

24-year-old male programmer,
Chief complaint: Paroxysmal dull pain around the navel for 2 days, accompanied by pasty diarrhea once daily.
No chronic disease history, allergy history, surgical history, or family genetic history.
Lifestyle habits: Prefers spicy diet, low exercise frequency (once weekly), no smoking or alcohol consumption.
No recent clear triggers, stress/emotional changes, or cold exposure.
Tongue examination shows thick white coating (pale red tongue body, rough surface with granular texture).
Pattern attributed to possible Spleen-Stomach Dampness Stagnation and impaired transportation-transformation function.

Tongue Image Description

Includes patient's tongue image:

and diagnostic content: According to the provided image and tongue coating classification results, the tongue coating presents as a thick, white, moss-like substance covering the tongue surface. The tongue body appears pale red with obvious white coating deposits.

Tongue Coating Characteristics:

- **Tongue Coating Color**: Primarily white.
- **Tongue Coating Thickness**: Thick.
- **Tongue Body Color**: Pale red.
- **Other Features**: The tongue coating surface is relatively rough, with a slight granular sensation.

Possible Health Issues:

Thick white coating in Traditional Chinese Medicine (TCM) is often associated with internal dampness accumulation, poor digestive system function, and improper diet. Specific aspects may involve:

1. **Digestive System Issues**: Possible indigestion, gastrointestinal dysfunction.
2. **Internal Dampness Pathogen**: Significant internal dampness, potentially related to diet, lifestyle, or environment.
3. **Phlegm-Dampness Stagnation**: Excessive phlegm-dampness affecting qi-blood circulation.
4. **Decreased Immunity**: Possibly related to reduced body immunity.

Please note: This is a preliminary analysis based solely on visual features of the tongue coating and cannot replace a professional doctor's diagnosis. If you experience discomfort or health concerns, seek timely medical consultation for detailed examination and diagnosis by a professional.

Diagnosis and Pattern Differentiation Conclusion

1. Summary of Chief Symptoms

- **Core Symptoms**: Paroxysmal dull periumbilical pain, pasty diarrhea (once daily)
- **Onset Characteristics**: Pain occurs paroxysmally, diarrhea is pasty rather than watery.
- **Aggravating Factors**: Long-term preference for spicy diet (easily generates damp-heat, referencing "dietary irregularities leading to Spleen qi deficiency and dysfunction of transportation" in *Diagnostic Standards for Spleen Deficiency Dampness Obstruction Pattern*), low exercise frequency (once weekly causing qi stagnation and damp pathogen congealment).
- **Accompanying Symptoms**: Although unmentioned by the patient, consistent with typical manifestations of Spleen Deficiency Dampness Encumbrance Pattern including: Reduced appetite (anorexia), abdominal distension (qi stagnation), and body heaviness (damp pathogen obstructing the body surface) (Referenced from knowledge graph search for "reduced eating, distension, body heaviness" symptoms in Spleen Deficiency Dampness Encumbrance Pattern).

2. Key Points of TCM Pattern Differentiation

- **Tongue Characteristics**: Thick white coating (pale red tongue body, rough surface with granular texture).
- **In-depth Analysis of Tongue and Pathomechanism**:
  - Thick white coating reflects the core pathomechanism of "weakened Spleen-Stomach transportation-transformation function, internal dampness encumbrance" (quoting unstructured data source).
  - White coating signifies internal cold-dampness retention; thick coating indicates exuberant damp-turbidity.
  - Rough granular texture suggests sticky stagnation of damp pathogen, qi stagnation.
  - Pale red tongue body indicates dampness predominating over deficiency, Spleen yang not yet deficient to the level of decline.
- **Corresponding Pulse**: Should correspond to "Soggy and Moderate Pulse" (referenced from knowledge graph search for typical pulse in Spleen Deficiency Dampness Encumbrance Pattern). Soggy pulse indicates dampness; moderate pulse indicates Spleen deficiency, aligning with signs of dampness encumbering the middle burner, Spleen qi insufficiency.

3. Diagnostic Conclusion

- **TCM Diagnosis**: Diarrhea (Spleen Deficiency Dampness Encumbrance Pattern).
- **Western Medicine Diagnosis**: Acute Intestinal Dysfunction.
- **Pathophysiological Mechanism**: Corresponds to dysfunction of autonomic nervous system regulation in the digestive system, abnormal intestinal hypermotility accompanied by digestive fluid secretion imbalance, leading to abdominal pain and pasty stools; related to long-term dietary irregularities causing intestinal flora dysbiosis and damage to the intestinal mucosal barrier function.

4. Diagnostic Basis

- From Knowledge Graph Search: Spleen Deficiency Dampness Encumbrance Pattern is defined as "Spleen qi deficiency and weakness, with internal retention of damp-turbidity, manifesting as reduced eating, distension, loose stools, body heaviness, pale enlarged tongueWhite moist or greasy coating, soggy moderate pulse, and other common symptoms" (direct quote of definition).
- From *Diagnostic Standards for Spleen Deficiency Dampness Obstruction Pattern*: Clearly states the etiology and pathomechanism of "dietary irregularities leading to Spleen qi deficiency, failure of transportation and transformation" (quoting standard document).
- From Web Search (Baidu Baike): Consistent with the core pathomechanism explanation: "Spleen governs the transportation and transformation of water-dampness, Spleen qi deficiency leads to internal water-dampness retention".

5. Treatment Principles

- **Fortify the Spleen and Transform Dampness**: Referencing the "Fortify the Spleen and Transform Dampness" principle in unstructured data, using modified Ginseng, Poria, and Atractylodes Macrocephala Powder (Shen Ling Bai Zhu San) (quoting treatment method for Spleen Deficiency Dampness Encumbrance Pattern from knowledge graph search).
- **Cautious Use of Cold/Cooling Herbs**: Adhering to the principle of "use cold-cooling herbs cautiously" (quoting unstructured data source), avoiding purgative substances like Rhubarb (Da Huang) or Mirabilite (Mang Xiao).
- **Harmonize Qi Mechanism**: Referencing "Harmonize Qi Mechanism" treatment principle (quoting unstructured data), combined with move qi and guide stagnation herbs.

6. Key Points for Differential Diagnosis

- Differentiation from Damp-Heat Diarrhea: This case has thick white coating (cold-dampness), not the yellow greasy coating (Huang Ni Tai) of damp-heat pattern; diarrhea is pasty, not the sudden gushing and urgent discharge with obvious foul odor of damp-heat pattern. Referencing the treatment note "cautious use of cold-cooling herbs" in unstructured data further confirms this case belongs to Spleen Deficiency Dampness Encumbrance Pattern, not damp-heat, avoiding mistaken use of heat-clearing and dampness-drying substances.

7. Prognosis Judgment

- Favorable prognosis. Referencing the "Fortify the Spleen and Transform Dampness" treatment principle in unstructured data, symptoms can be improved through spleen-fortifying, qi-supplementing, and aromatic dampness-transforming treatment.
- Combined with dietary advice to "avoid spicy, greasy, raw-cold foods" (quoting unstructured data source), preventing re-injury to Spleen-Stomach yang qi, can effectively prevent recurrence.

Basic Adjustment Plan for Spleen Deficiency Dampness Encumbrance Pattern

Dietary Adjustment Plan

According to web search data, patients with Spleen Deficiency Dampness Encumbrance Pattern should "Eat more spleen-fortifying and dampness-draining foods: such as Chinese yam (Shan Yao), lotus seed (Lian Zi), Poria (Fu Ling), Coix seed (Yi Yi Ren) (better stir-fried), hyacinth bean (Bai Bian Dou)" (from web search). Recommended specific recipes and cooking methods:

- **Spleen-Fortifying Dampness-Resolving Soup**: "Ingredients: Chinese yam 30g, lotus seed 30g, Poria 30g, Euryale seed (Qian Shi) 15g, appropriate amount of pork tripe or ribs. Method: Stew ingredients for 1 hour, add salt to taste, 2-3 times per week" (from web search).
- **Artemisia Scoparia, Hyacinth Bean, Poria Congee**: "Method: Wash Artemisia Scoparia (Yin Chen), decoct for broth, discard dregs; add washed hyacinth beans, Poria, and polished round-grained rice (Jing Mi), cook together; when congee is nearly done, add a little sugar and cook briefly" (from web search).
- **Millet Pumpkin Congee**: "Ingredients: Millet (Xiao Mi) 50g" (from web search), recommended to add appropriate amount of pumpkin cubes when cooking, benefits spleen-stomach harmonizing.
- **Soup-Based Regulation**: "If symptoms like poor appetite, gastric stuffiness, or diarrhea (signs of Spleen Deficiency Dampness Encumbrance) appear", use soup method, choose lean pork or chicken stewed with spleen-fortifying herbs.
- **Coix Seed Red Bean Soup**: Stew Coix seed (Yi Yi Ren) and red beans (Chi Xiao Dou) in proportion, benefits spleen-fortifying and dampness-resolving.

Cooking methods focus on stewing, boiling, and steaming; avoid frying and grilling; keep food light and easily digestible.

Dietary Schedule Tailored for Programmer Work Characteristics:

- Fixed Meal Times: Strictly follow breakfast 7-8 AM, lunch 12-1 PM, dinner 6-7 PM schedule.

Lifestyle and Behavior Adjustment

Addressing low exercise frequency, according to web search data, "Exercise is the most effective way for health preservation, benefits spleen-fortifying and dampness-resolving" (from web search). Specific exercise methods:

- Frequency: 5 times per week, 30-45 minutes each session, avoid within 1 hour after meals.
- Intensity: Moderate intensity, heart rate controlled within (220 - age) × 60%-70%.
- Time Allocation: Practice Ba Duan Jin (Eight-Section Brocade) for 15 minutes upon waking; walk for 30 minutes in the evening.
- Intensity Progression: Start with 3 times weekly initially, increase frequency by 1 session every two weeks.
- Post-Exercise Monitoring: Slight sweating without fatigue indicates appropriate intensity.

Specific TCM Exercise Methods:

- Ba Duan Jin (Eight-Section Brocade): Focus on "Regulating the Spleen and Stomach by Raising One Arm" and "Separating Heaven and Earth to Regulate the Triple Burner" forms. Practice 15-20 minutes daily.
- Tai Chi: Simplified 24-form, emphasize waist-abdomen rotation to promote spleen-stomach transportation.
- Walking Therapy: Slow walk for 30 minutes after dinner, combined with deep breathing; pace should induce slight sweating.
- Abdominal Massage: Clockwise massage 100 times, counter-clockwise 50 times daily morning and night, centered on the navel.

Specific Adjustment Methods for Programmers' Sedentary Work:

- Posture Adjustment: Use ergonomic chair, keep back straight, feet flat on floor.
- Timed Activity: Get up and move for 5 minutes every 45 minutes of work. Perform:
  - Neck Rotation: Slowly rotate neck, 10 times each side.
  - Shoulder Stretch: Cross hands and raise upwards, stretch backwards.
  - Waist Rotation: Hands on hips, rotate clockwise and counter-clockwise 10 times each.
- Desktop Regulation: Keep warm water at desk; avoid cold drinks.
- Eye Care: Look into distance for 5 minutes hourly, combined with eye acupoint massage.
- Work Environment: Keep workspace ventilated and dry; avoid damp environments worsening damp pathogen.
- Post-Lunch Regulation: Avoid immediate return to work after lunch; walk 10-15 minutes to aid digestion.

Sleep Regulation Recommendations:

- Sleep Duration: Ensure sleep before 10 PM, wake at 6 AM, guaranteeing 8 hours of sufficient sleep.
- Sleep Position: Recommend right side lying position to benefit spleen-stomach transportation, avoid abdominal pressure.
- Pre-Sleep Regulation: Soak feet in warm water (approx. 40°C) for 15 minutes 1 hour before sleep; can add a little ginger.
- Sleep Environment: Keep bedroom ventilated and dry, humidity controlled at 50%-60%, avoid dampness worsening damp pathogen.
- Pre-Sleep Restrictions: No eating within 3 hours of sleep; avoid excessive water intake.

Psychological Adjustment Suggestions

According to web search data, "Spleen deficiency individuals are prone to low mood; can try meditation, deep breathing to relieve stress" (from web search). Specific steps:

- Deep Breathing Practice: 3 times daily, 5 minutes each; use abdominal breathing (Inhale 4 sec, Hold 2 sec, Exhale 6 sec).

Figure 10: Report generated by MedMirror (Part I)

Quantification Standard for Warning Symptoms:

- Abdominal Pain Severity: Use 0-10 pain scale; score >5 requires medical attention.
- Diarrhea Frequency: Loose or watery stools >3 times daily is a warning signal.
- Weight Change: Weekly weight loss >1 kg warrants caution.
- Fatigue Level: Afternoon fatigue lasting >4 hours daily is abnormal.
- Appetite Assessment: Food intake reduced >50% for 3 consecutive days requires attention.
- Regulation Cycle: According to "Long-term regulation: Requires consolidation for 1-3 months after symptom relief" requirement (from web search), establish a 3-month assessment cycle.

The above adjustment plan should be implemented based on the patient's specific condition. Seek medical attention promptly if symptoms persist or worsen.

Related References

1. How to Regulate Spleen Deficiency Dampness Encumbrance Pattern - Baidu Health
2. What is Spleen Deficiency Dampness Encumbrance, What to Pay Attention To - Haodf Online
3. Treatment of Spleen Deficiency Dampness Encumbrance Pattern Eczema Based on "All Edema and Fullness Belong to the Spleen" Theory - Hanspub
4. How Long Does It Take to Regulate Spleen Deficiency Dampness Encumbrance Symptoms - Baidu Health - Medical Popularization
5. Spleen Deficiency Dampness Encumbrance Pattern - Spleen is the Pancreas - Zhihu Column

TCM Treatment Plan for Spleen Deficiency Dampness Encumbrance Pattern

(1) Herbal Prescription

Spleen-Fortifying Dampness-Resolving Core Formula

Modified Ginseng, Poria, and Atractylodes Macrocephala Powder (Shen Ling Bai Zhu San) (from knowledge graph search data). Base Formula: Codonopsis Root (Dang Shen) 15g,

- Severe diarrhea: Add Euryale Seed (Qian Shi) 12g,

- Pomegranate Rind (Shi Liu Pi) 9g.
- Poor appetite: Add Charred Three Immortals (Jiao San Xian) (Hawthorn, Medicated Leaven, Barley Sprout) 9g each.

Administration Method

Decoct 1 dose daily to obtain 300ml decoction. Take warm in two divided doses (morning and evening), 30 minutes after meals. 7 days constitute one course (from unstructured data referencing "modified Shen Ling Bai Zhu San" approach).

Shen Ling Bai Zhu San has the effects of "fortifying the spleen and supplementing qi, seeping dampness and stopping diarrhea" (from knowledge graph search data), suitable for diarrhea due to Spleen deficiency dampness excess. Formula Explanation: Codonopsis, Atractylodes Macrocephala, and Poria fortify the spleen and supplement qi as sovereign herbs; Chinese yam and lotus seed fortify the spleen and stop diarrhea, Hyacinth bean and Coix seed fortify the spleen and seep dampness as minister herbs; Villous Amomum Fruit aromatically awakens the spleen, Platycodon Root carries herbs upward and diffuses lung qi as assistant herbs; Licorice Root fortifies the spleen, harmonizes the middle, and moderates all herbs as envoy herb. The whole formula achieves the effect of fortifying the spleen, supplementing qi, seeping dampness, and stopping diarrhea.

Processing Methods: Atractylodes Macrocephala Rhizome should be stir-fried (Chao Bai Zhu) to enhance spleen-fortifying and diarrhea-stopping effects (from web search data); Coix seed is better stir-fried (Chao Yi Yi Ren) (from web search data); other herbs used in raw form. Decoction Key Points: Soak herbs for 30 minutes first. Add water to cover herbs by 2-3 cm. Boil vigorously, then simmer gently for 30 minutes. Add Villous Amomum Fruit later (add in last 5 minutes). Obtain 300ml decoction. Note: Avoid excessive use of cold-cooling herbs, as damp pathogen is sticky and stagnant; overuse can damage spleen yang, conversely worsening dampness obstruction (from unstructured data).

Contraindications: Dampness predominance contraindicates dampness-obstructing substances like Rehmannia (Di Huang) or Donkey-Hide Gelatin (E Jiao) (from unstructured data). Avoid raw-cold, greasy, sweet-cloying foods during medication. Adverse Reaction Handling: If poor appetite worsens or abdominal distension discomfort occurs, reduce dose or pause medication; discontinue immediately and seek medical attention for allergic reactions; increased bowel movements after taking medicine is a normal medicinal effect, adjust formula if persistent worsening.

Course Setting: 7 days constitute one course, generally requiring 2-3 courses. Efficacy Evaluation Standard: Based on improvement degree of common symptoms of Spleen Deficiency Dampness Encumbrance Pattern such as reduced eating, distension, loose stools, body heaviness (from knowledge graph search data). Categories: Markedly effective (symptoms disappear or significantly improve), Effective (symptoms partially improve), Ineffective (symptoms unchanged or worsen). Seek timely medical attention if severe abdominal pain, fever, or bloody stool occurs (from unstructured data).

(2) Acupuncture Plan

Main Points for Fortifying Spleen and Transforming Dampness

ST 36 Zusanli (bilat.), BL 20 Pishu (bilat.), SP 9 Yinlingquan (bilat.) (from web search data).

Auxiliary Points

CV 12 Zhongwan, ST 25 Tianshu, CV 6 Qihai; For severe diarrhea add BL 25 Dachangshu, ST 37 Shangjuxu; For abdominal distension add SP 4 Gongsun, PC 6 Neiguan.

Operation Key Points

Use 1.5-inch filiform needles. Apply even supplementation and drainage technique. Retain needles for 20-30 minutes after deqi, manipulating needles 1-2 times during retention. Perform once daily, 10 sessions constitute one course (from web search data on acupuncture methods for Zusanli and Pishu points).

Point Location: ST 36 Zusanli is located 3 cun below ST 35 Dubi, one finger-breadth lateral to the anterior crest of the tibia (from web search data); BL 20 Pishu is 1.5 cun lateral to the lower border of the T11 spinous process (from web search data). Needling Technique: Use 1.5-inch filiform needles. Insert perpendicularly 0.5-1.2 cun. Apply even supplementation and drainage technique to achieve deqi sensation. Retain needles 20-30 minutes, manipulating needles every 10 minutes during retention, aiming for local soreness/distension or radiating sensation.

Insertion Depth: ST 36 Zusanli 1-1.5 cun perpendicularly, SP 9 Yinlingquan 1-1.2 cun perpendicularly, BL 20 Pishu 0.5-0.8 cun obliquely towards the spine. Supplementation/Drainage Technique: Use even supplementation and drainage technique. After insertion and deqi, lift-thrust and twirl evenly at 60-90 times/minute. Manipulate every 10 minutes during retention. Focus on spleen-fortifying and qi-supplementing; avoid excessive drainage technique damaging spleen yang.

Needling Contraindications: Avoid blood vessels and important organs (from web search data). For Spleen Deficiency Dampness Encumbrance Pattern where tongue examination shows pale enlarged tongue, white moist or greasy coating, use drainage technique cautiously. Management of Abnormal Situations: For fainting during acupuncture, withdraw needles immediately and have patient lie flat (from web search data), administer warm sugar water; apply pressure to stop bleeding for hematoma; adjust depth for excessively strong needle sensation.

Treatment Frequency: Once daily or every other day. Course Arrangement: Acupuncture treatment typically constitutes 10 sessions per course (from web search data), with 2-3 days interval before next course (from web search data), consecutively for 2-3 courses. After symptom relief, reduce to twice weekly for maintenance treatment. Total treatment cycle 4-6 weeks.

(3) Tuina (Therapeutic Massage) Plan

Apply Spleen-Fortifying Dampness-Resolving Tuina techniques: ① Abdominal Rubbing (Mo Fu Fa): Rub abdomen clockwise centered on CV 8 Shenque for 5 minutes; ② Press-knead ST 36 Zusanli and SP 9 Yinlingquan for 2 minutes each; ③ Point-press BL 20 Pishu and BL 21 Weishu for 1 minute each; ④ Spine Pinching (Nie Ji Fa): Pinch spine from Changqiang (GV 1) to Dazhui (GV 14) 3-5 times. Perform once daily for 10 consecutive days (Referencing principle of exercise aiding qi transformation and dampness movement from unstructured data).

Operation Steps: ① Abdominal Rubbing: Rub abdomen clockwise centered on CV 8 Shenque (from web search data) gently for 5 minutes. ② Press-knead ST 36 Zusanli and SP 9 Yinlingquan: 2 minutes per point, aim for soreness/distension. ③ Point-press BL 20 Pishu and BL 21 Weishu: 1 minute per point, pressure gradually increasing. ④ Spine Pinching: Pinch spine from GV 1 to GV 14 3-5 times, taking about 3 minutes. Total operation time approx. 15 minutes, once daily.

Targeting common symptoms of Spleen Deficiency Dampness Encumbrance Pattern - abdominal distension with loose stools, body heaviness (from knowledge graph search data): ① For obvious distension: Focus press-kneading on CV 12 Zhongwan and ST 25 Tianshu, 2 minutes per point, add Separating Abdomen Yin-Yang technique. ② For severe loose stools: Extend abdominal rubbing to 8 minutes, add press-kneading CV 4 Guanyuan and CV 6 Qihai. ③ For body heaviness: Increase Bladder Channel (Back-Shu points) tuina, focus on BL 20 Pishu and BL 21 Weishu area, for 5 minutes.

4. Warnings to Monitor for Spleen Deficiency Dampness Encumbrance Pattern

Red Alert (Immediate Medical Attention)

- **Symptoms Present**: Severe unbearable abdominal pain, diarrhea exceeding 10 times daily with watery stools, high fever (temperature >38.5°C), obvious signs of dehydration (dry mouth/thirst, reduced urine output, poor skin turgor), blood or mucus in stool.
- **Management Plan**: Go to emergency department immediately. Perform blood tests, electrolytes, stool tests, etc. May require IV fluids and anti-infection treatment. Supplement TCM emergency measures: Acupuncture PC 6 Neiguan, ST 36 Zusanli, ST 25 Tianshu with even supplementation/drainage to harmonize center and stop diarrhea; Moxibustion CV 8 Shenque, CV 4 Guanyuan to rescue yang and secure collapse (for cold deficiency signs). Urgent tests: Blood culture, full electrolytes, liver/kidney function, ECG, abdominal CT. Key TCM Critical Pattern Identification: Any of the following indicates critical pattern: 1) Diarrhea gushing uncontrollably with cold limbs (Yang Qi Collapse); 2) Curled tongue, retracted scrotum, agitation, delirium (Damp-Toxin Sinking into Pericardium); 3) Faint pulse almost disappearing, oily sweat (Yin Exhaustion and Yang Collapse); 4) Skin withered, sunken eyes (Fluid-Qi Depletion). Immediately administer Shenfu Injection IV drip combined with moxibustion to rescue yang and secure collapse. Tongue Monitoring: Urgently monitor for coating peeling, dry cracked tongue surface, or ecchymosis. Pulse Monitoring: Focus on faint, thin, nearly disappearing pulse or excessively rapid/weak pulse (rate >120 bpm or <50 bpm). TCM Preventive Measures (Recovery Phase): Take Shen Ling Bai Zhu San preparation for 2-4 weeks to consolidate efficacy, combine with moxibustion CV 6 Qihai, CV 4 Guanyuan 3 times weekly. Follow-up Schedule: Regular follow-up on days 3, 7, 14, 30 post-discharge; weekly for first month, every 2 weeks for months 2-3.

Orange Alert (Seek Medical Attention within 24 Hours)

- **Symptoms Present**: Persistent unrelieved abdominal pain, diarrhea 5-10 times daily, mild dehydration (thirst, dark yellow urine), nausea/vomiting affecting eating, symptoms persisting 3 days without improvement.
- **Management Plan**: Visit gastroenterology department within 24 hours. TCM adjusts formula to enhance spleen-fortifying and dampness-resolving potency: Combine with symptomatic Western medicine if needed. Integrated TCM-WM Plan: TCM treatment with modified Dampness-Draining Stomach-Calming Decoction (Cang Zhu 10g, Hou Po 6g, Fu Ling 15g, Ze Xie 10g, Bai Zhu 10g, Gui Zhi 6g) to fortify spleen and drain dampness. WM Symptomatic Treatment: Oral rehydration solution to prevent dehydration, Smecta to stop diarrhea. Perform stool routine + occult blood if necessary. Additional Tests: CBC, Electrolytes, CRP, Abdominal Ultrasound. Tongue Monitoring: Focus on coating turning yellow-greasy, tongue body turning red. Pulse Monitoring: Beware pulse turning slippery-rapid (rate >90 bpm) or wiry-tight. TCM Preventive Measures: Add acupoint application (BL 20 Pishu, ST 36 Zusanli with Spleen-Fortifying Paste) twice weekly. Pattern Differentiation Key: Corresponds to "dampness stagnating transforming heat, qi mechanism reversal disorder" pathomechanism. Follow-up Schedule: Mandatory re-evaluation 24-48 hours post-treatment, then twice weekly until symptom relief.

Yellow Alert (Outpatient Follow-up Adjustment)

- **Symptoms Present**: Intermittent dull periumbilical pain, pasty stools 2-4 times daily, reduced appetite but able to eat, mild fatigue, persistent thick white rough tongue coating.
- **Management Plan**: Follow-up at outpatient clinic in 3-5 days. Adjust herbal formula (e.g., modified Shen Ling Bai Zhu San). Strengthen dietary control (avoid spicy/greasy). Increase light exercise. Specific Herbal Adjustment Suggestions: For patients with yellow alert symptoms, modify Shen Ling Bai Zhu San: - For obvious distension: Add Aucklandia Root (Mu Xiang) 6g, Villous Amomum Fruit (Sha Ren) 3g (decoct later) to rectify qi and transform dampness. - For persistent loose stools: Add Stir-fried Coix Seed (Chao Yi Yi Ren) 15g, Euryale Seed (Qian Shi) 10g to enhance spleen-fortifying and diarrhea-stopping. - For severe anorexia: Add Charred Three Immortals (Jiao San Xian) 10g each, Chicken Gizzard Lining (Ji Nei Jin) 6g to disperse food and guide stagnation. - For very thick greasy coating: Add Agastache Herb (Huo Xiang) 10g, Eupatorium Herb (Pei Lan) 10g to aromatically transform dampness. Diet & Exercise Quantification: Diet: Daily staple: Millet/Pumpkin or Yam Congee (300-400g), with boiled vegetables 200g, lean meat 50g. Strictly avoid spicy, greasy, raw-cold foods. Exercise: Slow walk 30 minutes after meals (pace 60-70 steps/minute), increase Ba Duan Jin practice by 1-2 sessions weekly (20 min/session). Basic Tests: Stool routine + occult blood, Tongue coating microbial test. Tongue Monitoring: Observe tongue body (pale? enlarged?), coating color (white? moist? greasy?), coating texture (rough? granular?) upon waking daily. Pulse Monitoring: Monitor if soggy moderate pulse (floating, thin, weak pulse; rate 60-70 bpm) remains stable. TCM Preventive Measures: Consume Spleen-

Figure 11: Report generated by MedMirror (Part II)